Vargha Fruzsina Sára

A nyelvi hasonlóság földrajzi mintázatai

Vargha Fruzsina Sára

A nyelvi hasonlóság földrajzi mintázatai

Magyar nyelvjárások dialektometriai elemzése

Magyar Nyelvtudományi Társaság
Budapest, 2017

A kutatás az OTKA PD 108442 számú pályázatának támogatásával készült. A könyv megírásakor a szerző Bolyai János Kutatási Ösztöndíjban részesült.

Lektorálta: Németh Miklós

A könyv megjelenését a Nemzeti Kulturális Alap támogatta.

ISBN 978-615-5061-17-2

Borítóterv: Krepler István, KHP Budapest

Tartalom

Ábrák, táblázatok és térképek jegyzéke

Ábrák

Táblázatok

Térképek

Bevezetés

Mint minden nyelv, a magyar is mutat területi különbségeket: egy-egy tájegység, település beszélőinek nyelvhasználatát más-más hangzásbeli, grammatikai vagy éppen szókészletbeli sajátságok jellemzik. A nyelv területi változatai a nyelvjárások, dialektusok. Az egyes nyelvjárások nem egymástól függetlenül léteznek; érintkeznek, hatnak egymásra, folyton változnak.

A nyelvjáráskutatás vagy dialektológia a nyelvi variabilitás (változatosság) vizsgálatának olyan területe, amely a térbeliséget állítja a középpontba. A dialektometria (nyelvjárásmérés) a dialektológia, azon belül is a nyelvföldrajz egyik részterülete. A nyelvföldrajz az egyes nyelvi jelenségek, nyelvi változók esetében megmutatkozó térbeli variabilitást kutatja és ábrázolja, jellemzően térképes formában. A dialektometria kvantitatív módon, statisztikai eszközökkel vizsgálja a nyelvjárások közötti hasonlóságot, illetve különbözőséget, az elemzések eredményeit többnyire térképen szemléltetve.

A dialektometriai kutatások elsősorban a nyelvatlaszok településekhez kötött adatait hasznosítják. A nyelvatlaszok a nyelvföldrajzi kutatások klasszikus produktumai, egyúttal a nyelvjáráskutatás alapvető forrásai. Az első nyelvföldrajzi kutatás Georg Wenker nevéhez fűződik, aki 1876-ban standard német mondatok helyi dialektusokra való „átfordítására" kért tanítókat, és válaszaik alapján kívánta térképen ábrázolni a mondatokban szereplő nyelvi jelenségek területi megoszlását. A nyelvföldrajzi kutatás és nyelvatlaszkészítés módszertanának kialakításában azonban sokkal inkább a romanista Jules Gilliéron munkássága volt meghatározó. Gilliéron 639 franciaországi településen végzett adatgyűjtést, 700 adatközlő bevonásával, kérdőív alapján. A gyűjtést egyetlen terepmunkás, Edmond Edmont végezte, aki kitűnő hallásának köszönhetően megbízhatóan tudta írásban visszaadni a kérdéseire kapott válaszokat. Biciklivel járta be Franciaországot, és folyamatosan postázta a kérdőfüzeteket a kutatócsoport számára. A jól szervezett munkamódszernek köszönhetően az adatgyűjtés 1896-os elindulásától mindössze 14 évre volt szükség, hogy 1910-ben az utolsó térképlap is napvilágot lásson. (A nyelvföldrajzi kutatások indulásáról és módszertanáról összefoglalóan lásd Chambers–Trudgill 1998: 13–31, Juhász 2001a.)

A későbbi európai, így a magyar nyelvjárási atlaszok is általában Gilliéron módszerét alapul véve készültek. A gyűjtés során elsősorban kisebb falvakat kerestek föl, ott is igyekeztek néhány gondosan kiválasztott, kevésbé iskolázott adatközlőt találni, általában férfiakat, mert az ő nyelvhasználatukat konzervatívabbnak tartották (l. Chambers–Trudgill 1998: 29–30). A nyelvatlaszok célja az volt, hogy tükrözzék egy-egy nyelv esetében a jellemző területi különbségeket, vagyis egy-egy kutatópont viszonylatában az adott településre jellemző nyelvjárási „normát" (a nyelvjárási normáról l. Kiss 2001a: 213–216). Legnagyobb adattárunk, A magyar nyelvjárások atlasza több szempontból is újítónak mondható, és nem csak azért, mert a kerékpár helyett gyakran az ötvenes évek fekete állami autói szállították a terepmunkásokat (összhangban a több évtizedes lemaradást felszámolni hivatott, nagyszabású tudománypolitikai tervekkel). Az újdonság abban állt, hogy a gyűjtők kutatópontonként több adatközlővel dolgoztak, nagyobb arányban kérdeztek meg nőket is, és fiatalabbakat is bevontak a vizsgálatba (Lőrincze 1975), sőt, az ellenőrző gyűjtések során a kor technikai színvonalát képviselő Mambo típusú magnetofonnal hangfelvételeket is készítettek.

A nyelvjárási gyűjtőmunka eredményének hagyományos közlési formája a papírra nyomtatott nyelvatlasz. A nyelvatlaszokban hagyományosan a kutatópontokat feltüntetve, az adott kutatópont mellé írják egy-egy szónak az adott településre jellemző vál-

tozatát. Egy-egy nagyobb területet lefedő nyelvatlasz több száz (esetleg több ezer) térképlapot is tartalmazhat, így több százezer nyelvi adatot ad közre. A magyar nyelvjárások atlasza informatizált változatában például 565 330 adat van. Az adatok nagy száma miatt ezek az adattárak különösen alkalmasak kvantitatív elemzési módszerekkel való feldolgozásra. A nyelvatlaszok kvantitatív vizsgálata eredeti, papír formájukban igen időigényes (noha korántsem lehetetlen, l. Imre 1971). Az adatok informatizálása és megfelelő kutatási módszerek fejlesztése révén hatékonyan végezhetünk akár olyan összetett elemzéseket is, amelyek papír-ceruza módszerrel nem valósíthatók meg. A nyelvatlaszkészítők generációjának örököseként a dialektometria művelői – ellentétben a nyelvi változatosság más kutatóival – alapvető feladatuknak tekintik a nyelvatlaszadatok feldolgozását (Nerbonne–Kretzschmar 2013: 2), a fáradságos munkával összegyűjtött adatok százezreinek valorizálását.

A XXI. század nyelvatlaszai (vagy nyelvjárási korpuszai) a számítógépes eljárásoknak köszönhetően újabb térképezési és elemzési technikák alkalmazásával készülnek. Már nemcsak kutatóponthoz kötötten, hanem az adatközlők paraméterei alapján is térképezhetjük a válaszokat (l. Bodó 2007a, Vargha 2011, P. Lakatos et al. 2012). A 2006-ban megjelent, Észak-Amerika nyelvjárási tagolódását ábrázoló atlasz (Labov–Ash–Boberg) pedig nem egyes szavak, kifejezések lejegyzéseit mutatja meg térképeken, hanem az egyes magánhangzók akusztikai paramétereit: az első és a második formáns értékét. Ráadásul maga a gyűjtés sem hagyományos terepmunka során, hanem telefonon történt, ami jelentősen növelte a kezdeti munkaszakasz hatékonyságát, biztosítva, hogy az első és az utolsó interjú elkészítése között ne teljen el túl hosszú idő. Szintén Észak-Amerikában végzett nyelvföldrajzi vizsgálatokat Jack Grieve (2016), informatikai módszerekkel, az interneten elérhető, amerikai napilapokban közölt olvasói levelekből összeállított, 38 millió szövegszavas írott nyelvi korpusz alapján, az egyes szavak, kifejezések területenként eltérő gyakoriságát is alapul véve. Az adatgyűjtés gyors és sok résztvevőt elérő legújabb formája a mobiltelefonos célalkalmazások fejlesztése és használata, egyelőre svájci és angliai példákkal (Leemann et al. 2016). Noha technikailag egyre messzebb kerülünk Edmond Edmont biciklijétől és ceruzájától, a nyelvi sokszínűség és változás dokumentálása az új évezredben is folytatódik.

A nyelvjárások kutatásán belül a kvantitatív megközelítésre, és különösen a dialektometriára jellemző leginkább a természettudományokban meghonosodott matematikai eljárások alkalmazása. A bonyolultabb műveletek, statisztikai elemzések elvégzéséhez elengedhetetlenek a megfelelő számítógépes alkalmazások, így a dialektometria módszertani szempontból a számítógépes dialektológia része. A magyar nyelvatlaszok dialektometriai kutatását, a nyelvjárások közötti hasonlóság automatikus adatösszevetésekkel történő vizsgálatát 2008-ban kezdtük el, az első eredményeket 2009-ben mutattuk be a Magyar Nyelvtudományi Társaság felolvasóülésén (Vargha–Vékás 2009). Az automatikus elemzéseknek nyilvánvaló előfeltétele, hogy legyenek megfelelő formában rögzített, elemezhető adataink. A magyar számítógépes dialektológiában az így rögzített, sokféleképpen felhasználható, egymással integrálható adatokat informatizált adatoknak nevezzük, magát a folyamatot pedig, ahogyan ezek az adatok létrejönnek, informatizálásnak (a magyar számítógépes dialektológia alapvetéseiről l. Vékás 2007). 2008-ban már százezres nagyságrendben álltak rendelkezésre ilyen adatok, elsősorban a magyar nyelvjárási atlaszok informatizálását támogató, hasonló módszerrel felépített, azonos nyelvészeti technológiát (Vékás 2007) alkalmazó projekteknek köszönhetően. A dialektometriai kutatás esetében így elsősorban megfelelő, a magyar adattárak és hangjelölési rendszer tulajdonságait figyelembe vevő

elemzési módszerek kidolgozására volt szükség. Jó példa tehát a dialektometriai elemzések esete arra, hogy az adatok informatizálásának időigényes munkafolyamata a későbbiekben, az adatok újrahasznosításával, sokszorosan is megtérülhet.

Jelen kötetben a magyar nyelvatlaszok közül négy adattár elemzéséről lesz szó, amelyek a következők: A magyar nyelvjárások atlasza (MNyA.), A romániai magyar nyelvjárások atlasza (RMNyA.), a Somogy-zalai nyelvatlasz (S–ZA.) és A moldvai csángó nyelvjárás atlasza (MCsNyA.). Az egyes atlaszokra a továbbiakban az itt szereplő rövidítések utalnak.

A nyelvatlaszok adatai jellemzően a huszadik század második felének rurális nyelvhasználatát tükrözik, és leginkább hangtani, illetve lexikai változatosságot mutatnak. Morfológiai jelenségek csak kis számban, a más nyelvi szinteken meglévő változatosság egyáltalán nem szerepel a térképeken. A nyelvi hasonlóság földrajzi mintázatai esetünkben csak az említett adattárak korlátain belül értelmezhetők.

A kötet öt fejezetből és az azokat lezáró összegzésből áll. Az 1. fejezet a dialektometria alapvető elméleti és módszertani kérdéseit ismerteti részletesen, magyar példákon szemléltetve az egyes elemzési módszereket. A 2. és 3. fejezet témája a fonetikai információ különböző mértékének hatása a dialektometriai elemzés eredményére, megmutatva egyúttal azt is, hogyan változnak a nyelvi hasonlóság földrajzi mintázatai a különböző nyelvi szintek függvényében. A 4. fejezet a nyelvatlaszok integrált elemzéséről szól, a két alapvető dialektológiai módszer – az adatok csoportosításán alapuló salzburgi és az automatikus adatelemzéssel működő groningeni – alkalmazásának tanulságaival a MNyA. és a RMNyA. integrált korpuszán. Ez a fejezet foglalkozik a magyar nyelvjárások és a köznyelv közti hasonlósággal is. A klaszteranalízis és a többdimenziós skálázás felhasználhatóságát a magyar nyelvjárások felosztásában, nyelvjárásterületek meghatározásában és a nyelvi kontinuum megjelenítésében az 5. fejezet mutatja be részletesen, a regionális atlaszok elemzésétől a teljes magyar nyelvterületen – vagyis a MNyA. és a RMNyA. integrált térképlapjain – végzett térképes kimutatásokig.

1. Dialektometria

Ez a fejezet a dialektometria alapvető célkitűzéseivel és módszertanával ismerteti meg az olvasót. Célja, hogy átfogó képet adjon a mára klasszikussá vált kutatási módszerekről, illetve a vizsgálatok menetéről. A következő kérdésekre keressük tehát a választ: Mire jó a dialektometria? Mire van szükségünk ahhoz, hogy dialektometriai kutatásokat végezhessünk? Hogyan illeszkednek a különböző módszertani irányok a magyar számítógépes dialektológiai kutatásokhoz?

1.1. Nyelvjáráshatárok, nyelvi hasonlóság és kvantitatív adatelemzés

A dialektológiai kutatások fontos célkitűzése már a kezdetektől, hogy a nyelvi jelenségek térbeli megoszlását vizsgálva magukat a nyelvjárásokat is csoportosítsák, nyelvjáráshatárokat állapítva meg, illetve nyelvjárásterületeket körülhatárolva (Juhász 2001b). Már a magyar nyelvjárásokról készült első nagyszabású elemzés, Balassa József munkája is részletesen, kartográfiai szempontból is precízen szemlélteti a magyar nyelvjárásoknak az adatelemzés eredményeképpen kirajzolódó területi egységeit (1891).

A nyelvjáráshatárok megállapításának klasszikus módszere nyelvjárási térképek elemzésén, jelenséghatárok megállapításán és az azokat jelző izoglosszák megrajzolásán alapul. Az egybeeső izoglosszák (izoglosszanyalábok) teszik lehetővé nyelvjárásterületek elhatárolását (Kiss 2001b: 72–74). Ezzel a módszerrel azonban csupán néhány, a területek elhatárolására a kutató által előzetesen alkalmasnak ítélt térképlap figyelembevételével van lehetőségünk, nem beszélve arról, hogy a legtöbb esetben a vizsgált jelenségek egyes változatainak előfordulása nem ad ki pontosan körülhatárolható területi egységeket.

Az izoglosszák megrajzolásánál igyekezett biztosabb módszert keresni nyelvjárási törésvonalak megállapítására Jean Séguy, aki elsőként alkalmazta a „dialektometria" kifejezést szomszédos[1] kutatópontok nyelvi hasonlóságának kvantitatív elemzésére. Séguy módszere értelmében ott feltételezhetünk nyelvjáráshatárt, ahol a szomszédos kutatópontok között nagyobb az adatok közti eltérés, mint más szomszédos kutatópontpárok esetében (Séguy 1973).

A kutatópontok közti nyelvi hasonlóságon alapuló elemzéseken kívül léteznek más, szintén kvantitatív, a térbeliség szempontját középpontba állító elemzések is. A magyar nyelvjárási adattárak alapján számos ilyen jellegű vizsgálat készült már. Példaként említhetjük a hangstatisztikai térképeket. Ilyen térképeket, a MNyA. alapján, elsőként Imre Samu készített (1971). Számítógépes dialektológiai eljárásokkal hasonló kimutatások előzetes kutatói klasszifikáció nélkül is elkészíthetők a MNyA. teljes kutatóponthálózatán, illetve akár integrált infomatizált adattárak esetében is (lásd pl. Bodó–Vargha 2008, Juhász 2011, Vargha 2007a, Vargha 2013). Vannak azonban olyan jelenségek is, amelyeknek a kvantitatív vizsgálata mindenképpen előzetes kutatói klasszifikációt igényel, hiszen automatikus számbavételük informatizált korpuszban sem lehetséges. Ilyen kimutatást készített Bodó Csanád a MCsNyA.-ban előforduló román kölcsönszók területiségéről (Bodó 2007b), megfelelő kódokkal kiegészítve az adattárat. Nyelvi változók kódolására és térképezésére nemcsak informatizált nyelvatlaszok, hanem területi szempontú szöveges adatbázisok felhasználásával is van lehetőség. Utóbbira példa a MNyA. szövegfelvételei alapján tett kísérlet a suksükölés

[1] Séguy, megfelelő eszközök hiányában, kizárólag a szomszédos kutatópontok adatait vetette össze, ami nyilvánvalóan nem elégséges a térbeli összefüggések megvilágítására.

hatvanas évekbeli területi elterjedtségének, terjedési dinamikájának vizsgálatára a Dunántúlon (Vargha 2007b), illetve az *l*-kiesés vizsgálata Vöő István hanganyaggal szinkronizált lejegyzéseiben (Fazakas 2013).

A dialektometria tágabb értelemben véve a kvantitatív nyelvföldrajzi kutatások sorába illeszthető. Lényeges különbség azonban a fentebb említett kvantitatív vizsgálatok és a dialektometriai elemzés között, hogy míg a többi kutatás egy-egy jelenség előfordulásának területi különbségeire fókuszál egy adott korpuszban, addig a dialektometriai kutatásokban az egyes jelenségek tekintetében megmutatkozó hasonlóság mértéke összeadódik, s így egyetlen, aggregált adatként mutatja meg a kutatópontpárok közti általános nyelvi hasonlóságot, illetve különbözőséget.

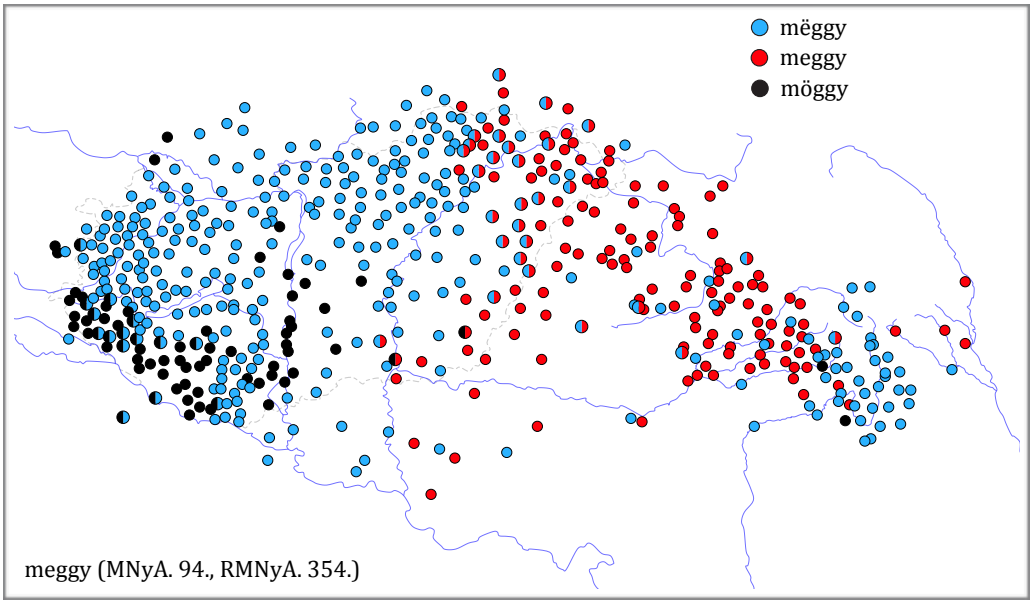
1.2. A salzburgi iskola

A dialektometria módszertanának kidolgozása Hans Goebl, klasszikus humán műveltségű, ám a matematikai eljárásokra is fogékony salzburgi romanista nyelvész nevéhez fűződik. Módszere az adatok térképenkénti osztályozásán, kutatói csoportosításán alapul. Az adatokat közlő eredeti térképek alapján a kutatók úgynevezett munkatérképeket hoznak létre, egy-egy jelenség szempontjából csoportosítva az adatokat. Az adatok munkatérképeken kialakított csoportosítását táblázatos formában összegzik, majd ebből az úgynevezett adatmátrixból hozzák létre a kutatópontpárok nyelvi hasonlósági értékeit megmutató hasonlósági mátrixot (a módszerről összefoglalóan l. Goebl 2010, 2011).

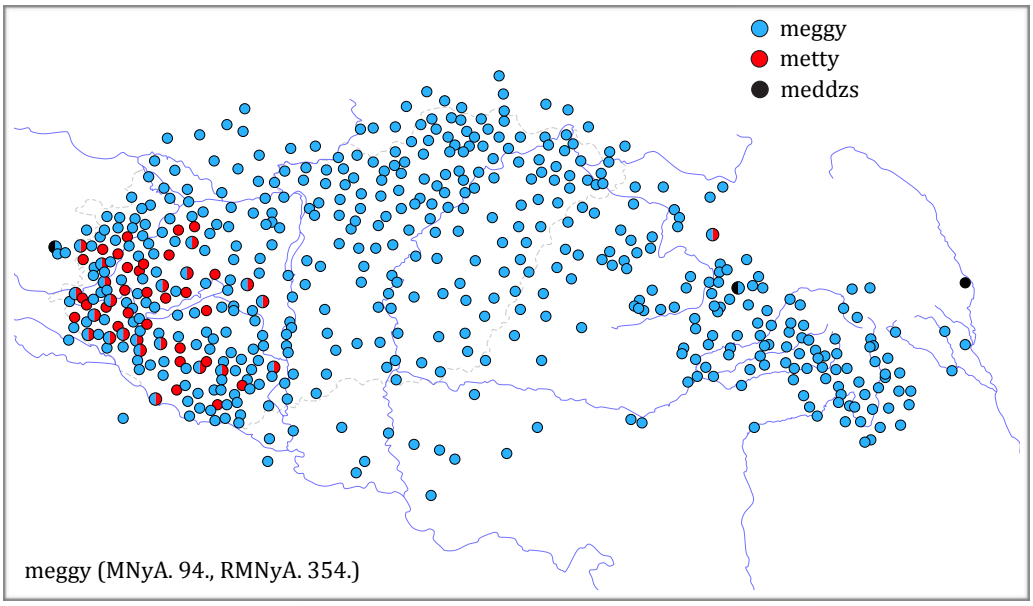
A munkatérképeket, illetve a kutatói csoportosítást úgy képzelhetjük el, mint egy egyszerű dialektológiai térképet, amelyen a térképlap adatait előre meghatározott szempont szerint csoportosítjuk, az egyes csoportokhoz különböző színeket vagy szimbólumokat rendelve. A MNyA. és a RMNyA. integrált térképlapjain 2015-ös kutatásunkban végeztünk dialektometriai elemzéseket Goebl módszerét alapul véve (Kocsis–Vargha 2016). Összesen 127-127 térképlap adatait elemeztük, 245 munkatérképet hozva létre. Az adatok csoportosítását az 1., 2., 3. és 4. térkép szemlélteti.

Az 1.1. és az 1.2. térkép a *meggy* fonetikai változatainak térbeli megoszlását szemlélteti a magánhangzó, illetve a szóvég mássalhangzó realizációja szempontjából. Az adatok csoportosításánál az alapjelet vettük figyelembe, így például a mellékjel nélküli és a nyíltabbnak jelölt zárt *ë* hangokat azonosnak vettük (természetesen más, a mellékjeleket is figyelembe vevő csoportosítás is lehetséges). Mindkét térképen tehát egyetlen szempont szerint csoportosítjuk az adatokat, vagyis egy térkép egyszerre egy nyelvi változó különböző variánsainak területi megoszlását mutatja meg.

Természetesen nem csak fonetikai szempontú csoportosításra van lehetőségünk, az 1.3. térkép az *egres* változatai területi megoszlásának szemléltetésével lexikai szempontú csoportosításra hoz példát. Ugyanakkor az adatok további klasszifikálásával lehetőségünk van az egyes lexikai csoportokon belül az adatok fonetikai szempontú csoportosítására. Az 1.4. térképlap például az 1.3. térképen világoskékkel jelölt *piszke, biszke, püszke, büszke* adatokat csoportosítja újra, a szókezdő mássalhangzó minősége szerint. Az utóbbi esetben maga a csoportosított térkép nyilván csak azokra a kutatópontokra értelmezhető, ahol a kérdéses lexikai változat előfordul. Az, hogy az adott csoportosítási szempont némely kutatópont adatára nem értelmezhető, bármely térképlap estében előfordulhat.

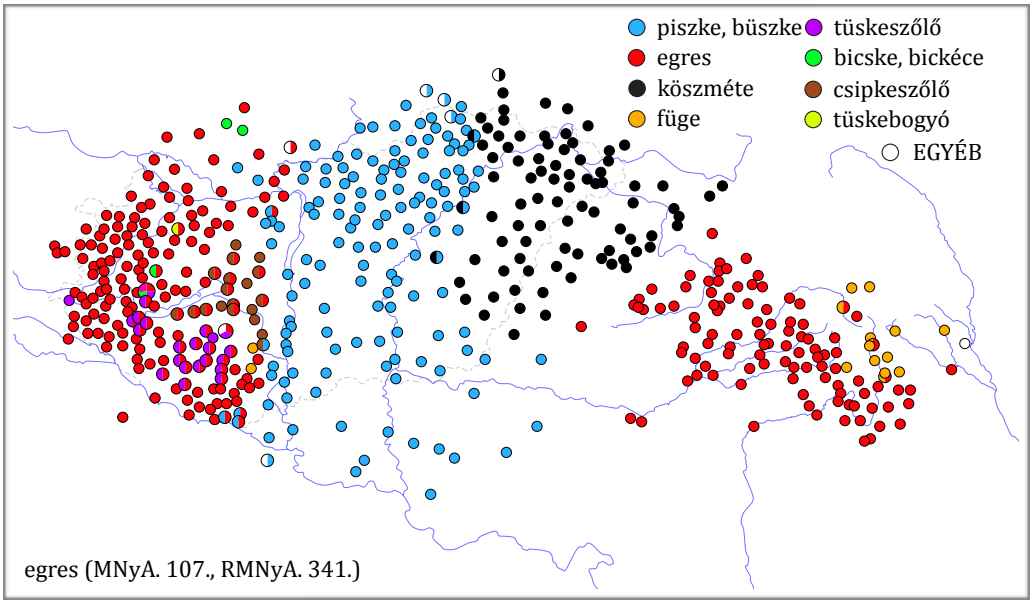


1.1. térkép: A *meggy* adatok csoportosítása a magánhangzó minősége szerint a MNyA. és a RMNyA. integrált térképén[2]

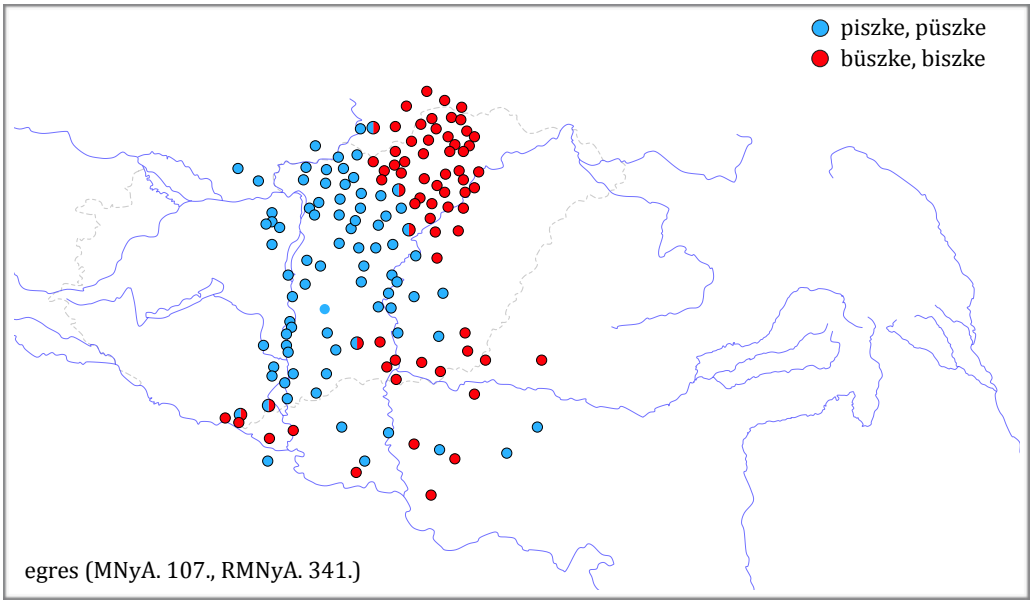


1.2. térkép: A *meggy* adatok csoportosítása a szótagzáró mássalhangzó minősége szerint a MNyA. és a RMNyA. integrált térkép

[2] Az itt látható térképek, további csoportosított munkatérképekkel együtt, elérhetők a MNyA. és a RMNyA. integrált dialektometriai elemzését bemutató honlapon: http://bihalbocs.hu/mnyarmnya/intterk.html.

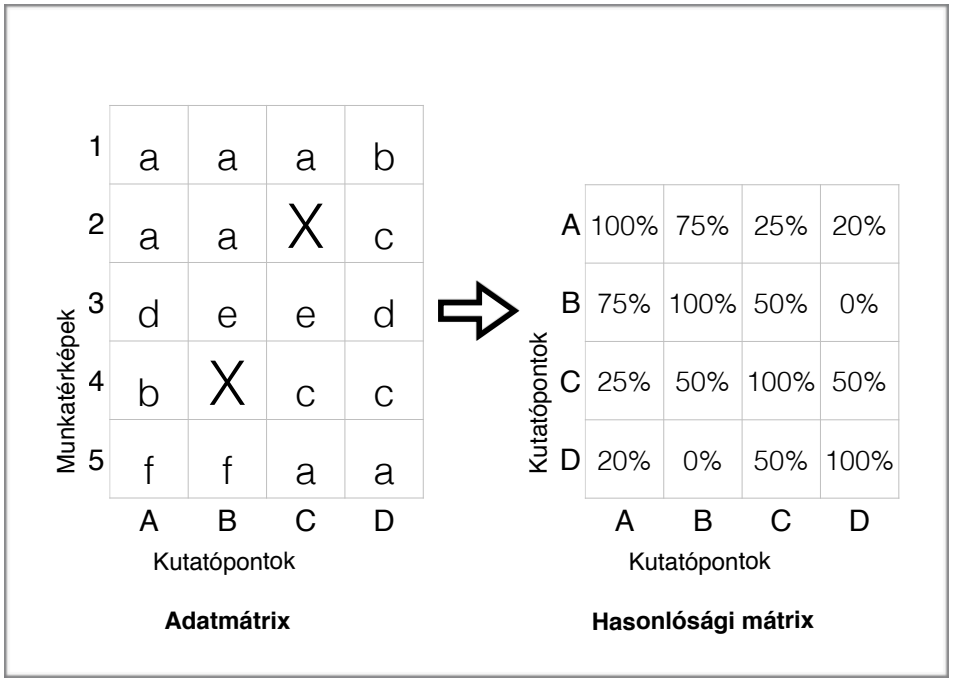


1.3. térkép: Az *egres* lexikai változatai a MNyA. és a RMNyA. integrált térképén



1.4. térkép: A *piszke* hangtani változatai a MNyA. és a RMNyA. integrált térképén

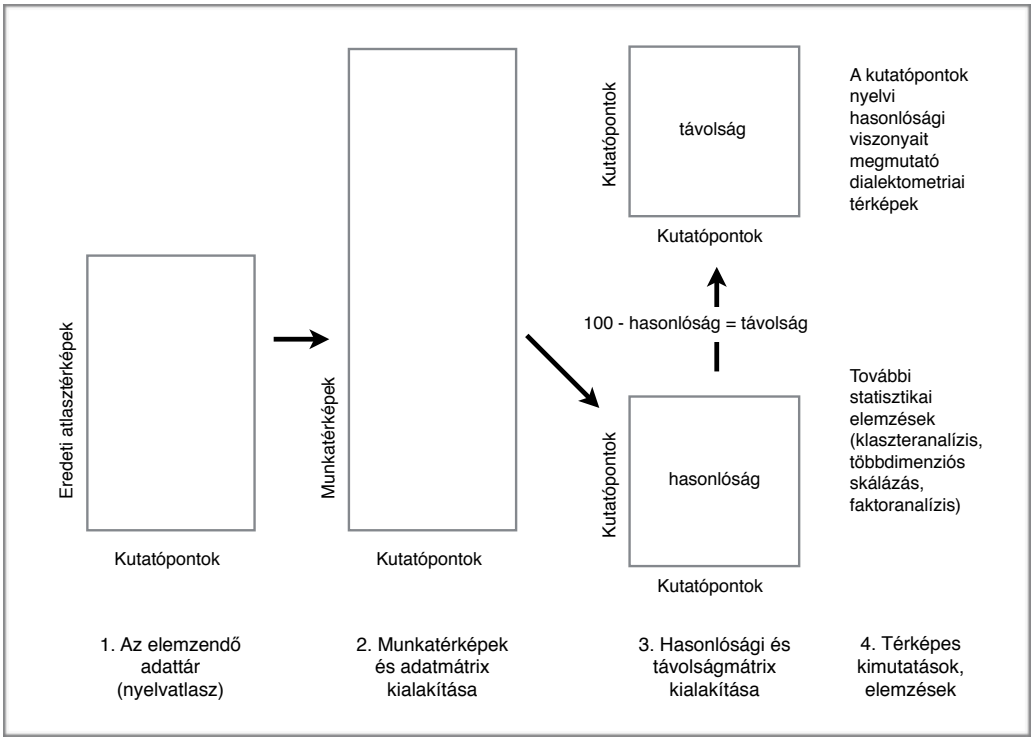
A munkatérképeken végzett csoportosítás eredményét Goebl adatmátrix formájában, összesítve mutatja be (Goebl 2010: 437). Az adatmátrix minden egyes munkatérkép esetében megmutatja, milyen csoportosító kódokat kaptak az adatok az egyes kutatópontokon a klasszifikáció során. A következő lépés a kutatópontok közti nyelvi hasonlóságot megmutató mátrix kialakítása az adatmátrix alapján. A hasonlósági mátrix minden egyes kutatópontpár esetében tartalmazza azt az értéket, amely a csoportosítás alapján megmutatja, milyen arányban egyeznek meg (kerülnek azonos csoportba) a két kutatópont adatai. Az 1.1. ábra a munkatérképekből kialakított adatmátrixot és az abból kialakított nyelvi hasonlósági mátrixot szemlélteti. Az itt látható minta öt munkatérképpel és négy kutatóponttal számol, a hasonlósági mátrixban szereplő érték két kutatópont viszonylatában annak függvénye, milyen arányban mutatnak egyezést az adatmátrixban szereplő csoportosító kódjaik. Az értékek így 100% (maximális egyezés) és 0% (maximális különbözőség) között lehetnek. Öt térképlap elemzése a valóságban nagyon kevés volna ahhoz, hogy releváns nyelvi hasonlósági mutatókhoz jussunk. A fent idézett, magyar anyagon végzett vizsgálatban például 245 munkatérképen csoportosítottuk az adatokat, Goebl Franciaország nyelvatlaszának (Atlas linguistique de la France) dialektometriai vizsgálatához 626 eredeti atlasztérkép alapján 1687 munkatérképet hozott létre (Goebl 2002).



1.1. ábra: A nyelvjárási térképek adatainak osztályozása, nyelvi hasonlósági mátrix kialakítása a salzburgi módszer szerint (Goebl 2010: 437 alapján)

A dialektometriai elemzés teljes folyamatát a salzburgi módszer szerint az 1.2. ábra szemlélteti. Az első lépés az elemzendő térképlapok korpuszának kijelölése. Ez jelentheti magát a teljes nyelvatlaszt, illetve annak egy részét. Minél több térképlapot vonunk be az elemzésbe, várhatóan annál pontosabb képet kaphatunk a kutatópontok közti nyelvi hasonlóságról. Az eredeti atlasztérképek alapján hozhatjuk létre az egy-egy elemzési szempontot érvényesítő munkatérképeket. A munkatérképek száma akár a

többszöröse is lehet az eredeti atlasztérképekének, hiszen a legtöbb térképlap adatai több kategorizációs szempont mentén is csoportosíthatók.



1.2. ábra: A dialektometriai elemzés folyamatábrája a salzburgi módszer szerint (Goebl 2006: 413 alapján)

Ezt követően (ahogyan azt már az 1.1. ábrán is láttuk) kerül sor az adatmátrix alapján a kutatópontpárok közötti nyelvi hasonlósági értékek kiszámítására, vagyis a nyelvi hasonlósági mátrix kialakítására. A mátrix megmutatja, hogy minden egyes kutatópont mely kutatópontokkal mutat nagyobb, melyekkel kisebb hasonlóságot. A munkatérképek alapján nem csak a nyelvi hasonlóság, hanem a nyelvi különbözőség is kiszámítható. Bármelyik mátrixot hozzuk létre elsőként, abból a másik már egyenesen következik, hiszen a különbözőség és a hasonlóság összege mindig 100%. Így a hasonlóságból kiindulva a nyelvi távolságot úgy kapjuk meg, hogy a 100%-ból egyszerűen kivonjuk a hasonlóság mértékét.

A nyelvi hasonlóságot (vagy különbözőséget) megmutató mátrixokat interaktív térképre is vetíthetjük, amely egy színskála mentén megmutatja, hogy egy kiválasztott kutatópont mely kutatópontokkal mutat nagyobb, melyekkel kisebb hasonlóságot. A lehetőségeink ezzel még nem érnek véget, hiszen a kész mátrixokon további statisztikai vizsgálatokat végezhetünk a kutatópontok közti nyelvi hasonlóság alapján, amelyeket szintén akár térképen is szemléltethetünk (l. részletesen az 1.5. alfejezetben).

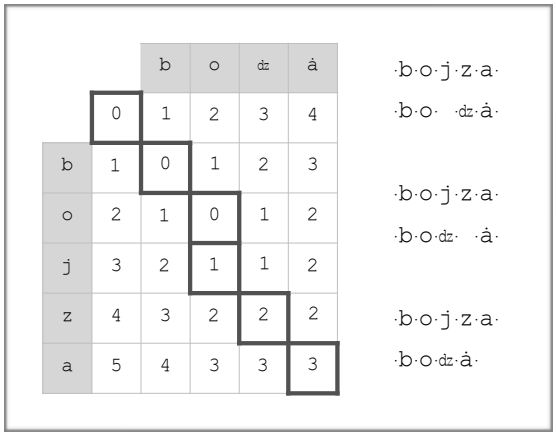
A salzburgi módszer alkalmazása az utóbbi években könnyebben adaptálhatóvá vált más kutatóhelyek számára is (lásd pl. Montemagni 2008, Perea 2010) a kifejlesztett, általánosan használható szoftverek révén. Az első ilyen, kifejezetten az eredeti, Goebl-féle elemzési módszert alkalmazó program a Visual DialectoMetry (VDM), amelyet Edgar Haimerl fejlesztett ki Salzburgban (részletesebben l. Goebl 2006), de újabban vannak más, online elérhető alkalmazások is. Ilyen a DiaTech internetes alkal-

mazás, amely a VDM programba beépített lehetőségeken túl további elemzési módszereket kínál a felhasználó számára, illetve Goebl módszeréhez képest előnye, hogy kutatópontonként minden munkatérképen több adattal is számolhatunk (Aurrekoetxea et al. 2016). Szintén szabadon letölthető a GeoLing alkalmazás, amelyet eredetileg a Sprachatlas von Bayerisch-Schwaben informatizálására és dialektometriai elemzésére fejlesztettek ki (https://www.geoling.net), de más adattárak fogadására és elemzésére is alkalmas.

1.3. Adatpárok automatikus összevetése

Ahogyan az eddigiekből kitűnik, a salzburgi dialektometriai módszer kulcsa az adatok kutatói csoportosítása. A csoportosítás azonban igen időigényes, így meglehetősen költséges is. Több matematikai módszer is rendelkezésünkre áll azonban arra, hogy az adatainkat automatikusan vessük össze egymással egy számítógépes programba épített algoritmus segítségével (az alkalmazható eljárásokról l. részletesebben Navarro 2001). Az egyik ilyen, a helyesírásellenőrzésben és a természettudományok számos területén is alkalmazott eljárást Vladimir Iosifovich Levenshtein orosz matematikus dolgozta ki, 1965-ben. Az alábbiakban az adatok automatikus összevetésén alapuló Levenshtein-féle eljárás dialektometriai alkalmazásával ismerkedünk meg.

Kessler (1995) használta elsőként dialektometriai kutatásban a Levenshtein-algoritmust, az ír nyelvjárások elemzésére. A módszer lényege, hogy az adatok összevetése nem előzetes kutatói klasszifikáció alapján, hanem a lejegyzett adatok elemzésével, automatikusan történik. A módszer további tesztelése és általános használatának megalapozása a dialektometriai kutatásokban a groningeni egyetem kutatói, elsősorban John Nerbonne és Wilbert Heeringa nevéhez fűződik (Heeringa 2004; Heeringa & Nerbonne 2001, 2013; Nerbonne et al 1996). Az algoritmus két, fonetikusan lejegyzett és betűláncnak felfogott adat egymáshoz képesti távolságát méri, kiszámítva, hogyan lehet legkönnyebben (legkevesebb lépéssel, legkisebb transzformációs „költséggel") „átforgatni" az egyik betűláncot a másikba (azaz a köztük lévő különbözőség mértékét fejezi ki számszerűsített formában). Két, betűláncként felfogott nyelvi adat automatikus összevetését az 1.3. ábra szemlélteti.



1.3. ábra: Betűláncok összevetése a Levenshtein-algoritmus használatával

A betűláncok összevetésekor valójában azt vizsgáljuk, mennyi változtatással forgatható át legegyszerűbben az egyik szó a másikba betoldással, törléssel és cserével. Az 1.3. ábra az összevetés mechanizmusát mutatja meg két konkrét nyelvi adat, a *bodzá* és a

bojza esetében. A két szóban a betűk (karakterek) száma megegyezik ugyan, de a nyelvjárási adatoknál a hangokat vetjük egymással össze, a *dz* betűkapcsolatot tehát egyetlen (rövid) hangnak vesszük, az egyezményes hangjelölés gyakorlatának megfelelően. A *bojza* egy hanggal hosszabb, ezért mindenképp szükségünk van egy betoldásra. A *bodzá* változatban szereplő hangok közül csak kettő található meg a *bojza*-ban, ezért szükséges még két cserét is elvégeznünk. Ez összesen három művelet. A műveletek számát a hosszabb szó elemeinek számához viszonyítva megkapjuk a két szó közötti különbség mértékét százalékban kifejezve. Esetünkben ez 60%. Vagyis a két szó 60%-ban különböző, így értelemszerűen 40%-ban egyezik, hiszen a nyelvi távolság és hasonlóság összesen mindig 100%-ot ad ki: a két összevetett adat bizonyos mértékben egyezést, bizonyos mértékben eltérést mutat. Az átalakítás lehetséges útjai közül mindig azzal számolunk, amelyiknek a legkisebb a költsége, vagyis a legkevesebb művelettel jár. A 40%-os egyezés csak más adatösszevetésekhez képest értelmezhető: például a *bojza* és a *borza* nagyobb arányban, 80%-ban mutat egyezést.

A Levenshtein-algoritmus eredeti változatában mindig 1-nek számolja minden szükséges változtatás transzformációs költségét. Levenshtein az eredeti algoritmushoz képest javasolt egy olyan alternatívát is, ahol a cserék költsége nem egy, hanem 2. Nerbonne és munkatársai (1996) az algoritmusnak ezzel a változatával számolják ki az adatok közti távolságot. Lehetséges továbbá az eredeti algoritmus bővítése a hangátvetés lehetőségével, így két változat között (pl. *kalarábé ~ karalábé*) két lépés helyett egy lépésre csökkenthető a távolság (alkalmazására l. pl. Heeringa 2004).

Az alapjelek közti hasonlóságon túl további kérdés a mellékjelek kezelése az elemzés során. Már Kessler kutatásában (1995) fölmerült ez a kérdés. Kessler összetett szimbólumokkal dolgozott, vagyis az alapjel-mellékjel kombinációt önálló hangnak vette. Annak érdekében, hogy a kisebb fonetikai különbségek kevésbé módosítsák két hasonló, csak egy-két mellékjelben különböző adat távolságát, különböző súlyt adott a szükséges változtatásoknak a hangok artikulációs eltérésének függvényében. A különböző módszerekkel készült elemzések összevetése során azonban arra jutott, nincs számottevő előnye, ha van egyáltalán, a fonetikai különbségeket súlyozó elemzésnek. Nerbonne és munkatársai (1996) külön karakterként kezelik az alapjeleket és mellékjeleket a Reeks Nederlandse Dialectatlasse elemzésekor, de a mellékjeleket érintő változtatásoknak kisebb súlyt adnak az elemzés során. Heeringa elemzésében eltekint a mellékjelektől (2004: 65), viszont az egyes alapjelek közti különbségeket súlyozza (Heeringa 2004: 124–125). Levenshtein algoritmusának eredeti változatában bármely két hang közötti különbség azonos hatással van a két adat közti távolság kiszámítására, hiszen az eljárás során nem értelmezzük a fonetikai tartalmukat. Éppen olyan különbség van tehát egy *e* és egy *ä* között, mint egy *e* és egy *k* között. Lehetséges az algoritmus módosítása oly módon, hogy a különböző karakterpárok közti különbség mértékét súlyozzuk, illetve korlátozhatjuk, hogy melyik hang melyekkel cserélődhet ki, és melyekkel nem, annak függvényében, hangtanilag mennyire elképzelhető, hogy valamely nyelvjárási változat egyik hangja egy másik nyelvváltozatban éppen azzal a másik hanggal helyettesítődjön. A különböző, a Levenshtein-algoritmus módosításán alapuló elemzések nem hoznak számottevően más eredményt (Nerbonne et al 1996), sőt, Kessler összevetései alapján (1995) a karaktereket azonos súllyal tekintő, eredeti elemzésmód a korábbi dialektológiai kutatásokkal jobban egybevágó nyelvi mintázatokat rajzol ki.

További kérdés annak kezelése, ha egy kutatóponton több adat is előfordul. Két, több adatot tartalmazó kutatópont összevetésére Nerbonne és Kleiweg (2003) azt a módszert alkalmazza, hogy A kutatópont összes adatát (a_1, a_2, a_3… a_n) összeveti B ku-

tatópont összes adatával (b_1, b_2, b_3… b_n), és minden egyes adat esetében a legkisebb távolságot tartja meg, vagyis A kutatópont minden adatát azzal a B kutatóponton szereplő adattal vetjük össze, amelyikre a legjobban hasonlít. Egy adatpárral azonban csak egyszer számolunk az eljárás során. Aurrekoetxea és munkatársai (2013) amellett érvelnek, hogy az apróbb változtatásokra kevésbé érzékeny, biztosabb eredményhez jutunk, ha mindkét kutatópont irányából egyaránt elvégezve az összevetéseket kétszer számolunk a szimmetrikus adatpárokkal. A két kutatópont közti nyelvi távolság minden esetben az adatok közti legkisebb távolságok átlaga.

A Levenshtein-algoritmus használatán alapuló dialektomeria elsősorban fonetikai különbségek kimutatására használatos. Azt mutatja meg tehát, milyen mértékben hasonlítanak (térnek el egymástól) az egyes nyelvjárások fonetikailag. A MNyA. *árpa* térképlapján például eltérést találunk az egyes nyelvjárások között a magánhangzók minőségét illetően, pl. *árpa, árpo, ārpä, ᵃórpä*. Akkor azonban, ha az adatainkból automatikusan információt vonunk el, létrehozhatjuk az adatok reprezentációjának olyan szintjeit, ahol már nem érvényesülnek a fonetikai különbségek. Ha szélsőségesen leegyszerűsítjük a lejegyzést, tehát megfosztjuk a legtöbb fonetikai információtól, csak a legelemibb különbségeket tartva meg, akkor a hangtanilag kisebb eltéréseket mutató adataink egybeesnek, a nagyobb (jellemzően lexikai) különbségek viszont megmaradnak. Az *árpa* különböző hangtani változatai a lejegyzés egyszerűsítésével azonossá vagy közel azonossá válnak, s így a hasonlóságuk mértéke a 100%-hoz közelít, a ’denevér’ jelentésű lexikai változatok, pl. *szárnyasegér, bőrmadár* viszont radikális egyszerűsítés után is jelentős mértékben különbözni fognak egymástól (hasonlóságuk 0% közeli marad). A fentiek értelmében a lejegyzés fonetikai információtartalmát automatizált eljárással módosítva, az adatok előzetes (manuális) csoportosítása nélkül végezhetünk fonológiai pontosságú vagy lexikai jellegű dialektometriai összevetéseket a Levenshtein-algoritmus használatával (Vargha 2015).

A magyar nyelvjárási atlaszok dialektometriai elemzésekor térképről térképre haladva, páronként összevetjük minden kutatópont összes adatát valamennyi kutatópont összes adatával, az adatpárok hasonlósági értékeit pedig minden egyes kutatópontpár esetében átlagoljuk, így megtudjuk, hogy egy kutatópont minden más kutatóponttal rendre átlagosan milyen mértékű hasonlóságot mutat. Az összevetések számszerűsített végeredménye éppen úgy, ahogy a salzburgi módszer esetében, itt is egy hasonlósági mátrix, amely megmutatja, átlagosan milyen arányban mutatnak egyezést egymással az egyes kutatópontok adatai. Így bármelyik kutatópontról megállapíthatjuk, hogy adatai átlagosan mely kutatópontok adataival mutat nagyobb, és melyekkel kisebb hasonlóságot. (Saját magával – értelemszerűen – minden kutatópont itt is 100%-os hasonlóságot mutat.)

Fontos leszögeznünk, hogy a bármilyen módszerrel kiszámított nyelvi hasonlóság nem jelent feltétlenül „nyelvjárás-rokonságot”. Elképzelhető, hogy olyan kutatópontok közt is nagyobb egyezés mutatható ki az atlaszadatok alapján, amelyek nem állnak egymással településtörténeti kapcsolatban. Jó eszköz lehet azonban a dialektometria a rokonság gyanújának megerősítésére, a településtörténeti kapcsolatok pontosítására, például nyelvjárásszigetek eredetének vizsgálatakor.

Összevetve egymással az adatok csoportosításán alapuló salzburgi módszert és a Levenshtein algoritmusa segítségével végzett elemzéseket a következő megállapításokat tehetjük. A csoportosítás – annak ellenére, hogy a nyelvi adatok informatizálása nem feltétel – igen időigényes, csak tapasztalt dialektológus által végezhető, hiszen a kialakított csoportosításon múlik az elemzések végeredménye. Ráadásul a csoportosítás kialakítása során számos nehézségbe ütközhetünk, amelyek megnehezítik az

adatok világosan elkülönülő kategóriákba való besorolását. Mivel azonban minden munkatérképet egyetlen szempont érvényesítésével alakítunk ki, a jelenségtérképek könnyen sorolhatók különböző csoportokba, amelyeknek a hatása a nyelvi hasonlóság alakulására viszonylag egyszerűen összevethető, elemezhető (Goebl 2005, Pickl et al. 2014). Az adatok automatikus összevetésekor maga az elemzés megfelelő technológiával pillanatok alatt elvégezhető, ebben az esetben előfeltétel azonban, hogy legyen megfelelő mennyiségű informatizált adatunk. Az adatok rögzítése igen időigényes, vagyis költséges, megfelelő irányítás mellett azonban nincs szükség alaposabb dialektológiai szaktudásra az elvégzéséhez.

Adatok	Csoportosítás A	Csoportosítás B	Levenshtein
tüskę̂szöllőű	a	a	61,6%
csipkęszöllő	a	b	

Adatok	Csoportosítás 1.	Csoportosítás 2.	Levenshtein
mëtty	a	a	20,0%
mę̆ggy	b	b	

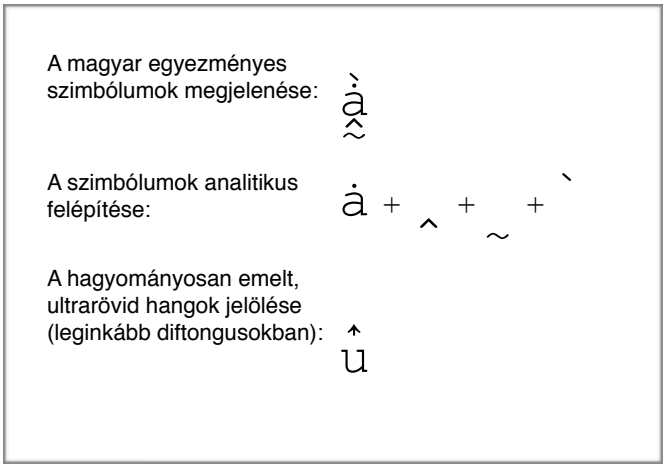
1.4. ábra: Adatpárok lehetséges csoportosítása és Levenshtein algoritmussal való összevetése

Az 1.4. ábra néhány, a fentebb látható munkatérképeken szereplő adat összevetését szemlélteti két különböző módszerrel. A *tüskeszőlő* és a *csipkeszőlő* lexikai változatok összevetésekor a Levenshtein algoritmussal végzett elemzés esetében azt kapjuk, hogy a két adat 61,6%-ban mutat egyezést. Ha munkatérképeken szeretnénk csoportosítani az *egres* különböző lexikai változatait, el kell döntenünk, hogy az adatokban rejlő azonosságot (*szöllő)* vagy különbözőséget *(tüske* vs. *csipke)* vesszük figyelembe. Az első (A) esetben azonos, a második (B) esetben különböző csoportosító kódot kap a két adat, tehát a kutatói döntés függvényében 100%, vagy 0% lesz a hasonlóság mértéke (kvantitatív megközelítés esetén elfogadható, ha esetenként nem egyértelmű az adatok csoportokba sorolása). A *meggy* hangtani változatainak esetében az adatokat Levenshtein-algoritmussal elemezve a köztük lévő hasonlóságot szintén egyetlen szám fejezi ki (20%). Kutatói csoportosítás esetén a magánhangzó, illetve a mássalhangzó képzésében lévő különbséget két külön munkatérképpel tudjuk megragadni (l. az 1.1. és az 1.2. térképet), mindkettő 0%-os hasonlóságot eredményez. A későbbi elemzések során, ha külön szeretnénk elemezni a magánhangzók és a mássalhangzók különbségéből kirajzolódó nyelvi hasonlósági mintázatokat, akkor a csoportosított térképek esetében, a munkatérképeket külön kategorizálva ezt könnyen megtehetjük. A különböző nyelvi szintek hatása a dialektometriai elemzésre a fonetikus lejegyzés különböző információtartalmú változatain keresztül Levenshtein algoritmussal készült elemzéssel automatikusan is megvalósítható (l. a 3. fejezetet).

1.4. Magyar számítógépes dialektológia, dialektometria

A magyar számítógépes dialektológia már húszéves múltra tekinthet vissza. Ha a MNyA. Balogh Lajos kezdeményezésére indult számítógépes rögzítését tekintjük az első lépésnek, amelynek eredményét sok évvel később beépítettük a már informatizált térképlapok adatbázisába, akkor 25 évvel is számolhatunk (Balogh–Kiss 1992). A magyar nyelvjárási adatok hatékony számítógépes rögzítéséhez és kezeléséhez szükséges nyelvészeti technológiák kidolgozója és sokáig egyedüli fejlesztője Vékás Domokos volt. A legkorszerűbb példákat és irányelveket alapul véve, azokat több vonatkozásban radikálisan továbbgondolva vágott bele a Bihalbocs nevű alkalmazás fejlesztésébe és a papír atlaszok számítógépes feldolgozásába, az informatizált változatok kialakításakor is maximálisan megőrizve a magyar egyezményes hangjelölés alapelveit, szem előtt tartva hagyományait is (a magyar számítógépes dialektológia indulásáról összefoglalóan l. Vékás 2007).

Vékás Domokos olyan kódrendszert dolgozott ki a magyar dialektológiai adatok informatizálására, amely messzemenően figyelembe veszi a kiindulási forrás vizuális megjelenését is (vagyis a magyar egyezményes hangjelölésben használatos alapjelek és mellékjelek grafikus formáit). A kialakított kódrendszer révén a Bihalbocsban mozaikszerűen lehet fölépíteni a hangokat (fonetikai szimbólumokat) az alapjelekből (pl. *ë, e, é, ȧ, ä*) és az azok ejtését módosító diakritikus jelekből (pl. a zártabb ejtést jelölő, fölfelé mutató ék vagy a nazális ejtést jelölő hullámvonal). A kereshetőséget és a lejegyzés fonetikai értelmezhetőségét a mellékjelek kötött sorrendje is biztosítja (l. az 1.5. ábrát).

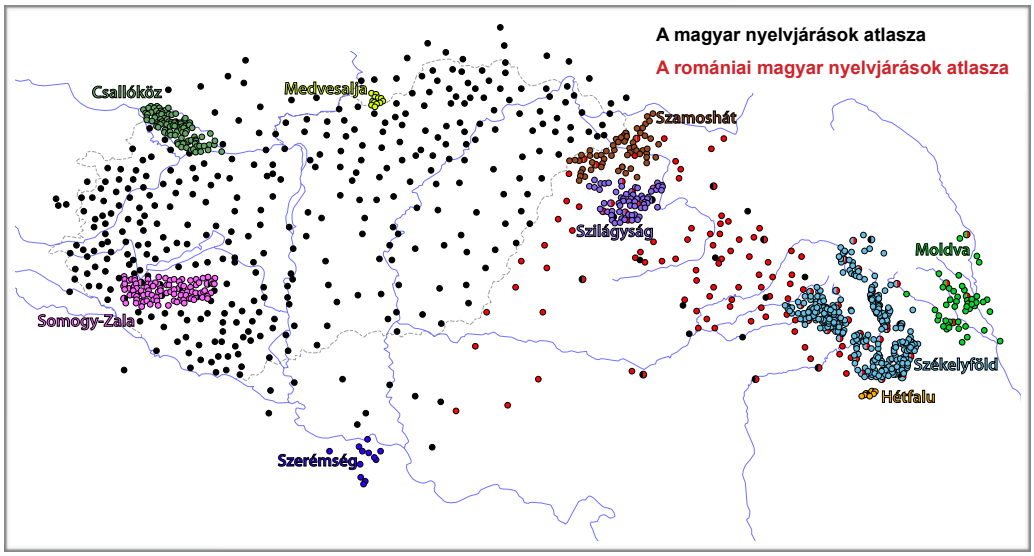


1.5. ábra: A magyar egyezményes alapjelek és mellékjelek megjelenése és felépítése a Bihalbocsban

A lejegyzés során minden alapjel és mellékjel könnyen begépelhető a Bihalbocs speciális billentyűkiosztásának köszönhetően. A mellékjelek helyes sorrendjét a program biztosítja. Mivel az adatok külalakja az informatizált változatban lényegében megegyezik a forrásban lévővel, már az adatbevitel során kisebb a hibázás esélye, illetve sokkal gyorsabb, hatékonyabb az ellenőrzés folyamata is. Az egyetlen formai különbség a magyar egyezményesben szokásos és a Bihalbocsban alkalmazott jelölésmód között az ultrarövid hangok ábrázolása, amelyeket hagyományosan emeléssel szoktak ábrázolni a lejegyzésben. Valójában azonban az ultrarövid ejtés jelölése funkciójában semmiben sem különbözik más, a hang alapértelmezett időtartamát módosító jelektől,

mint például a félhosszúság vagy a hosszú ejtés mellékjele. A Bihalbocsban, megtartva a hangjelölés koherenciáját, az ultrarövid ejtést is mellékjellel, a hang fölött elhelyezett kis nyíllal jelöljük, utalva így a hagyományos gyakorlatban, így az atlaszadatokban is alkalmazott emelésre is (l. 1.5. ábra).

A Bihalbocsban kidolgozott módszerrel már több mint egymillió atlaszadatot rögzítettünk, nagyobb részt pályázati támogatással, NKFP és OTKA finanszírozású projektekben, de több példa van arra is, hogy egyéni kutatómunka, szakdolgozat vagy doktori disszertáció elkészítéséhez vág bele valaki egy-egy nyelvatlasz teljes vagy részleges informatizálásába. A Bihalboccsal informatizált, illetve a munkafolyamat elejétől a Bihalboccsal készülő (ilyen a Csallóközi nyelvatlasz, a vállalkozásról l. részlesebben Menyhárt–Presinszky 2013) adattárak integrált kutatópont-hálózatát az 1.5. térkép szemlélteti. A következő fejezetekben a térképen szereplők közül a Bevezetőben is említett négy adattár dialektometriai elemzéséről lesz szó: Somogy–zalai nyelvatlasz (S–ZA.), A magyar nyelvjárások atlasza (MNyA.), A romániai magyar nyelvjárások atlasza (RMNyA.) és A moldvai csángó nyelvjárás atlasza (MCsNyA.). Az elemzésre kiválasztott nyelvatlaszok közül teljes a MNyA., a MCsNyA. és a S–ZA. informatizálása, a RMNyA. 3297 térképlapjának mintegy fele, összesen 1525 térképlapnyi áll rendelkezésre megfelelően rögzített formában.



1.5. térkép: Bihalboccsal informatizált adattárak integrált kutatópont-hálózata

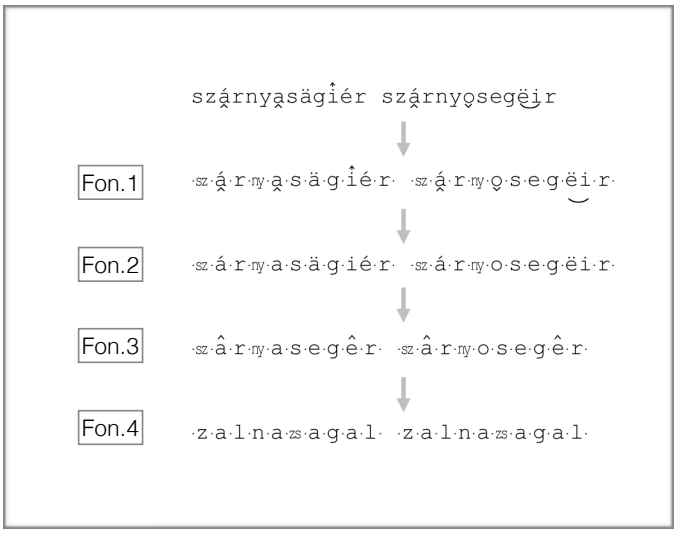
A magyar nyelvjárások dialektometriai elemzése elsősorban az informatizált atlaszadatok automatikus összevetésével történik, amelyhez a Levenshtein algoritmus eredeti változatát használjuk, vagyis minden változtatás (betoldás, törlés, csere) költsége egyaránt 1. Az elemzésben az alapjelek és a mellékjelek is külön karakternek számítanak, a lejegyzésben betűkombinációval ábrázolt mássalhangzókat (pl. *gy*) azonban egy hangnak, vagyis egyetlen karakternek feleltetjük meg (l. 1.6. ábra), a geminátaként jelölt mássalhangzók (pl. *ggy*) két elemből állnak. Mivel az alapjelek és mellékjelek fonetikai értelmezése a Bihalbocsban maximálisan biztosított, lehetőségünk van arra, hogy automatikus konverziók segítségével létrehozzuk az eredeti finom fonetikai lejegyzés más, egyszerűsített formáit (l. 1.6. ábra). A lejegyzés így létrehozott, fonetikailag kevésbé részletes változatai alkalmasak arra, hogy a fonetikai információ

mennyiségének a nyelvi hasonlósági viszonyok alakulására gyakorolt hatását vizsgáljuk, illetve az egyes nyelvi szintek (fonetikai, fonológiai, lexikai szint) szerint végezzünk elemzéseket (l. részletesebben a 2. és a 3. fejezetben). A lejegyzés egyszerűsítésének hasznossága fölmerül akkor is, ha két adattárat integrálunk, hiszen még ha valamennyi adattárban a magyar egyezményes lejegyzést alkalmazzák is, mégsem teljesen azonosan teszik. A különbségek leginkább a mellékjelek használatát érintik, illetve a diftongusok jelölésmódjának különbségeiben mutatkoznak meg pl. a MNyA. és a RMNyA. viszonylatában (l. Vargha 2015b, illetve a 4. fejezetet).

A diftongusokat az eredeti mellékjeles lejegyzés elemzésekor két karakterként vesszük figyelembe, tehát például egy *ó* és egy *óu* diftongus (utóbbinál az ultrarövid ejtést emelés helyett a karakter fölötti nyilacska, vagyis mellékjel fejezi ki) 33%-ban azonos, egy *ó* és egy *òu* azonban már maximálisan különbözik egymástól, vagyis a köztük lévő hasonlóság mértéke 0%. Az eredeti mellékjeles lejegyzés egyszerűsítésével a Bihalbocsban létrehozhatjuk az adatok más reprezentációs szintjeit különböző mennyiségű fonetikai információ megtartásával. Dönthetünk például úgy, hogy összevonjuk a hasonló magánhangzókat egymással, és eltekintük az időtartambeli különbségektől is, ebben az esetben az utóbbi példában szereplő két magánhangzó (*ó* és *òu*) már 50%-ban hasonlítani fog egymásra, hiszen az egyszerűsítés következtében *o* és *ou* lesz belőlük. Lehetőségünk van arra is, hogy a diftongusokat hosszú monoftongusoknak feleltessük meg, ebben az esetben az eredeti lejegyzésben *ó* és *òu*-ként lejegyzett hangok azonossá válnak, hiszen az *òu* diftongusból hosszú *ó* lesz.

A magyar nyelvatlaszokban, különösen a MNyA.-ban, teljesen általános, hogy egy kutatóponton több, esetenként akár tucatnyi adat is szerepel. Ilyenkor a hasonlóság mértékének megállapításához két kutatópont viszonylatában a következőképpen járunk el: összevetjük először az egyik (legyen A) kutatópont összes adatát (pl. *me̩ggy, mëggy, metty)* a másik kutatópont (legyen B) összes adatával (pl. *mëggy, möggy*), egyenként, az első adattól az utolsóig, mindig a legmagasabb hasonlósági értéket tartva meg. Összevetve A kutatópont adatait B kutatópont adataival a három alakváltozat esetében a hasonlóság mértéke a következőképpen alakul: a *me̩ggy* B kutatópont mindkét adatával 60%-ban mutat egyezést, a *mëggy* a *mëggy*-gyel 100%-ban, a *metty* a *mëggy* és a *möggy* adattal egyaránt 25%-ban azonos. A kutatópont hasonlósága B kutatóponthoz a három százalékos érték átlaga, vagyis 61,67%. Összevetve B kutatópont adatait A kutatópont adataival (*mëggy* vs. *mëggy* = 100%, *möggy* vs. *mëggy* = 75%) a hasonlóság értéke 87,5%. A hasonlóság mértékét tehát mindkét kutatópont irányából külön kiszámítjuk, végül pedig ezt a két értéket átlagoljuk, példánknak megfelelően így 74,6%-ot kapunk.

A Bihalbocsban az adattárak kutatóponthálózatait közös térképen is megjelenítheti, így tetszés szerint egyszerre több informatizált atlasszal is dolgozhatunk. Az azonos címszavú térképlapok integrálásával és egyidejű elemzésével készíthetünk a teljes nyelvterületet ábrázoló dialektometriai térképeket (l. 4. fejezet és 5.5. fejezet), illetve együttesen elemezhetjük például a RMNyA. és a MCsNyA. térképlapjait, tágabb összefüggésben vizsgálva a moldvai nyelvjárások nyelvi hasonlóságát a Székelyfölddel és a Mezőséggel (l. 5.4. fejezet).

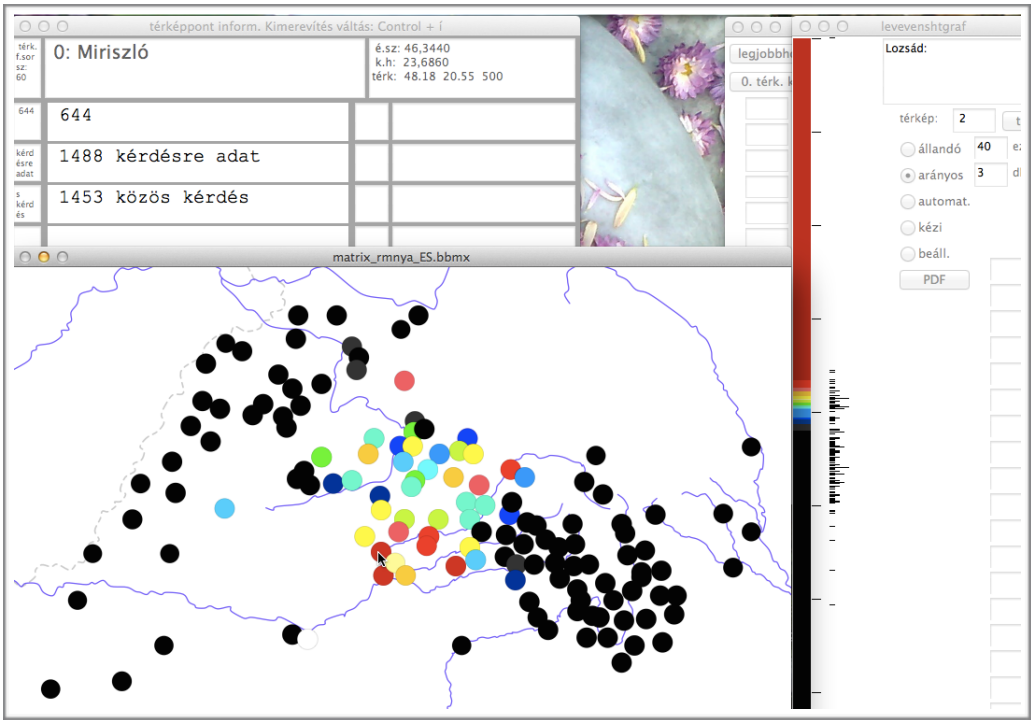


1.6. ábra: Az adatok automatikus egyszerűsítése a Bihalbocsban. Forrás: Vargha 2015

Az adatok automatikus összevetésén túl végeztünk (a salzburgi módszerhez hasonlóan) csoportosításon alapuló összevető elemzéseket is. Egyrészt a MNyA. és a RMNyA. integrált térképlapjain a Bihalbocsba beépített adatcsoportosítási funkció segítségével (l. Kocsis–Vargha 2016, valamint a 4. fejezetet), másrészt a S–ZA. adattárként közölt részén, Király Lajos csoportosítása alapján (l. Vékás–Vargha 2009, valamint a 2. fejezetet).

1.5. A hasonlósági mátrixok térképezése és statisztikai elemzése

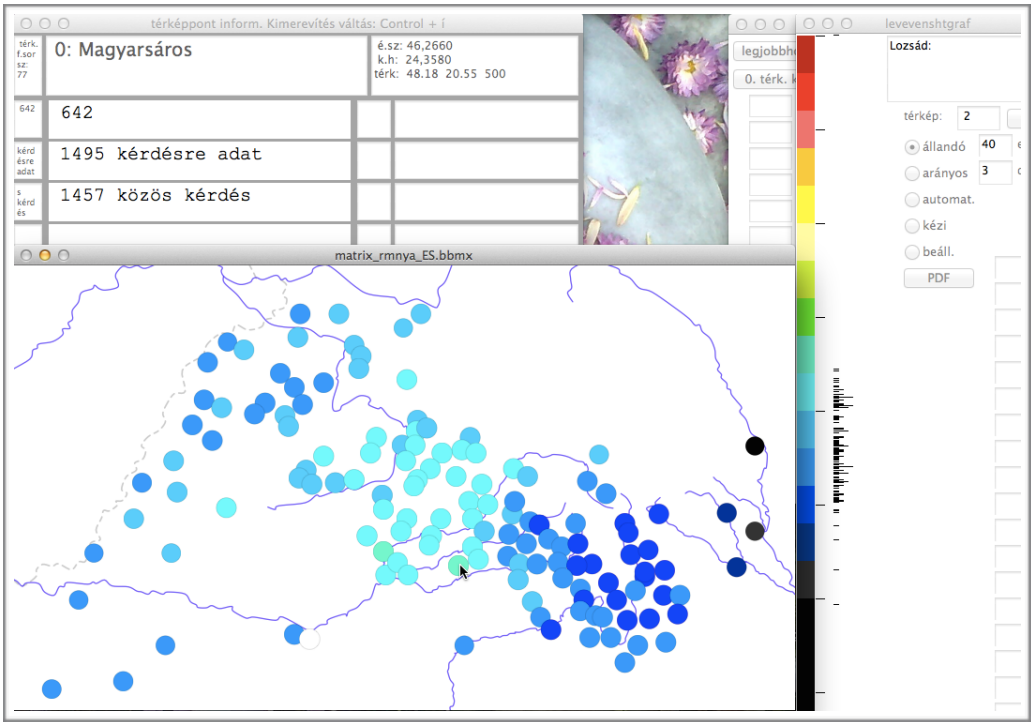
A dialektometriai elemzés végeredményét, vagyis a kutatópontok közti nyelvi hasonlóság (vagy távolság) mértékét megmutató mátrixot térképezhetjük is. Az interaktív térképen egy kutatópontot kiválasztva egy színskála mentén megnézhetjük, hogy mely kutatópontokkal mutat nagyobb, melyekkel kisebb hasonlóságot. A Bihalbocsban különböző beállítások révén megnézhetjük azt is, mely kutatópontok mutatnak a kiválasztott kutatóponttal leginkább hasonlóságot (l. 1.7. ábra), illetve azt is, általában milyen mértékben találtunk egyezést az adatokban a kiválasztott kutatópont és a többi kutatópont között (l. 1.8. ábra). A 1.7. és a 1.8. ábra az interaktív dialektometriai térkép működését mutatja a Bihalbocsban a RMNyA. kutatóponthálózatán. A kiválasztott kutatópont (Lozsád) fehér színben látszik, a többi kutatópont színe a színskála beállításainak megfelelően alakul. A nagyobb nyelvi hasonlóságot meleg színek (a sötét vöröstől indulva), a kisebb nyelvi hasonlóságot hideg színek (egészen a feketéig) jelölik. A kurzor mozgatásával a kutatópontok fölött a Térképpont információs ablakból megtudhatjuk, hogy egy-egy kutatópont milyen mértékben (az értéket ezrelékben adjuk meg) mutat hasonlóságot a kiválasztott kutatóponttal (esetünkben Lozsáddal), az elemzett térképlapok közül hány térképlapon találtunk adatot, és ezek közül mennyi volt közös a kiválasztott kutatóponttal, vagyis **mennyi** összevetés alapján kaptuk meg a két kutatópont közti hasonlóság mértékét kifejező százalékos (ezrelékes) értéket.



1.7. ábra: Interaktív dialektometriai térkép a Bihalbocsban. A kiválasztott kutatópont Lozsád, a kurzor a vele leginkább hasonlóságot mutató Miriszló fölött van.

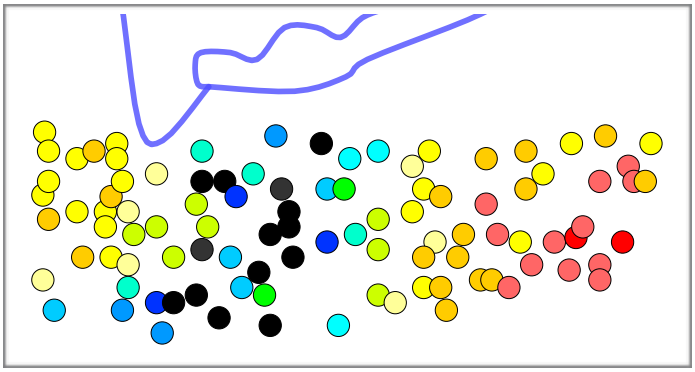
Az 1.7. ábrán, Lozsád dialektometriai térképén a beállításoknak megfelelően három kutatópontonként változnak a színek, a színskálának megfelelően (l. az ábra jobb oldalán). A színskálán az egyes színek által reprezentált intervallumot a színsávok hossza jelöli. A Lozsádhoz leginkább hasonló kutatópontokat jelző első intervallum (sötét vörös) azért annyira hosszú, mert a nyelvileg legközelebbi Miriszló 64,4%-ban mutat vele hasonlóságot, a harmadik legközelebbi kutatópont, Magyarkapud pedig 63,4%-ban, így a sötét vörös szín 99,9%-tól egészen 63,4%-ig tart (a kijelölt kutatópont saját magával értelemszerűen 100%-ban mutat hasonlóságot, azaz azonosságot, amit fehér színnel jelölünk). Mivel a beállítás értelmében három kutatópontonként színt váltunk, így viszonylag hamar elfogynak a lehetőségek a vörös és a fekete között (a RMNyA.-nak összesen 136 kutatópontja van), így a kutatópontok nagyrészt feketében látszanak; a térképen látható, hogy ez a szín fogja át a kutatópontok nagyobb részét.

Az 1.8. ábrán szintén Lozsád dialektometriai térképe látható, csak ebben az esetben a színskála színei mindig egyenlő intervallumot, 4%-ot (40‰-et) fednek le, 100%-tól kezdődően, ahogyan azt a színskála és a beállítások mutatják. A térképen nincsenek meleg színek, hiszen a legnagyobb mért hasonlóság is csak 64,4%. Az ábrán a kurzor Magyarsáros fölött van, itt találjuk a második legnagyobb értéket, 64,2%-ot. A Lozsádhoz leginkább hasonló két kutatópont (Miriszló és Magyarsáros) zöldes színben, a mezőségi kutatópontok türkiz színben látszanak, a legkevésbé hasonlóak a keleti székely, illetve a moldvai kutatópontok.



1.8. ábra: Interaktív dialektometriai térkép a Bihalbocsban. A kiválasztott kutatópont Lozsád, a kurzor a hozzá második leghasonlóbb kutatópont, Magyarsáros fölött van.

A példákból látszik, hogy egyetlen mátrix interaktív térképen való szemléltetésével is többféle lehetőségünk van a kutatópontok közti nyelvi hasonlóság megjelenítésére. Ugyanabból a korpuszból azonban különböző módszerekkel (csoportosítás, automatikus adatösszevetés) vagy a lejegyzés fonetikai tartalmának változtatásával különböző mátrixokat is készíthetünk, amelyeket aztán rendre térképezhetünk is. Így összevethetjük egymással ugyanazon kutatópont különböző elemzéseken alapuló dialektometriai térképeit, illetve készíthetünk összevető elemzéseket a különböző mátrixokról is. Akár a kutatópontok közti földrajzi távolság mátrixát is összevethetjük a nyelvjárási távolsággal. Az 1.6. térkép a S–ZA. finom fonetikai mátrixa és a kutatópontok közti földrajzi közelség korrelációját szemlélteti kutatópontonként. Azon kutatópontok esetében, ahol a földrajzi közelség és a nyelvjárási hasonlóság együttjárása nagyobb, melegebb a kutatópont színe (piros, narancs, sárga), ahol kisebb, ott hidegebb (zöld, kék), illetve szürke, fekete. Az atlasz keleti (külső-somogyi) kutatópontjai esetében, illetve a nyugati (zalai) részen a nyelvjárási hasonlóság a földrajzi közelség alapján jól megjósolható, ezzel szemben az atlasz középső részétől kissé nyugatra (Somogy megye nyugati részén) már jóval kevésbé, miközben a két mátrix közti korrelációs együttható értéke 0,757, vagyis a két mátrix összességében erős pozitív korrelációt mutat. A következő fejezetekben számos további példát láthatunk majd mátrixok közötti összevetésre.

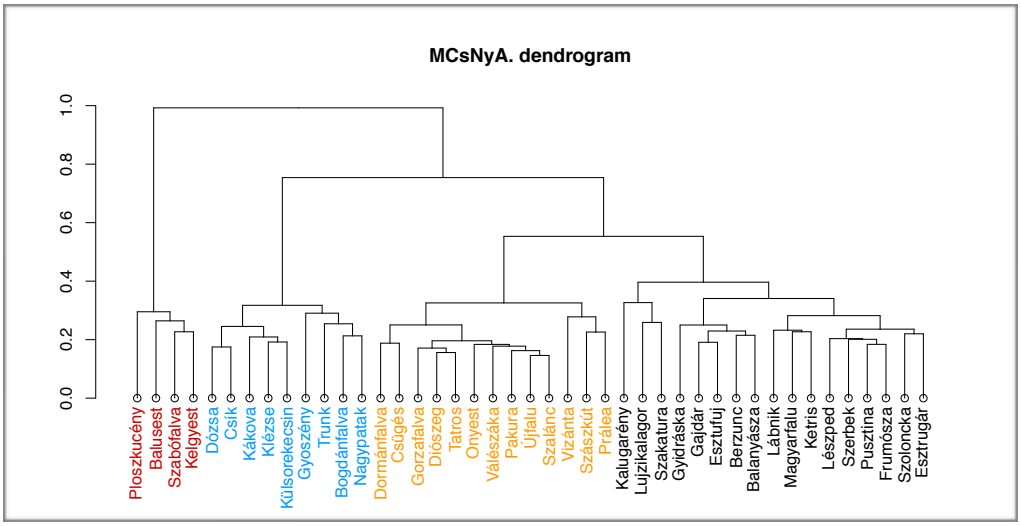


1.6. térkép: A nyelvi hasonlóság és a földrajzi közelség korrelációja a S–ZA.-ban

A mátrixunkat a térképezésen és más mátrixokkal való összevetésen túl további statisztikai elemzéseknek vethetjük alá. A legkorábban alkalmazott eljárás a klaszteranalízis, amely a nyelvjárások automatikus csoportosítását, nyelvjárásterületek dialektometriai alapú kialakítását célozza. Többféle klaszterezési eljárás is létezik, a nyelvjárások csoportosítására az eddigi módszertani kutatások alapján a Ward-féle klaszteranalízis tűnik a legjobbnak (Goebl 2011: 153–154, Grieve 2014: 68–69, Prokic–Nerbonne 2008). A klaszteranalízis lényege, hogy úgy hozzunk létre csoportokat, hogy az egy egységbe sorolt elemek (például egy nyelvatlasz kutatópontjai) minél inkább hasonlítsanak egymásra és minél inkább különbözzenek a más típusba sorolt elemektől. A Ward-féle eljárás matematikailag azzal biztosítja e követelmények (hasonlóság, különbözőség) minél jobb teljesülését, hogy a klaszterek összevonásakor a lehető legkisebb legyen a szórásnégyzet növekedése a létrejövő új klaszteren belül. Ezzel az eljárással viszonylag hasonló elemszámú csoportok jönnek létre, a felosztás más klaszterezési módszerekhez képest inkább mutat hasonlóságot a korábbi, kvalitatív elemzéseken alapuló területi felosztásokkal.

A MCsNyA. fonetikailag érzékeny mátrixa alapján készített klaszterelemzés dendrogramja látható az 1.9. ábrán. A dendrogram megmutatja, hogyan vonhatjuk össze csoportokba a kutatópontokat a köztük lévő nyelvi hasonlóság alapján. Különböző színek mutatják a kutatópontok felosztását négy csoport elkülönítésekor. Az ábráról leolvashatjuk azt is, hogyan alakíthatnánk ki kisebb alcsoportokat az elemzés alapján, illetve, hogy hogyan alakulna a kutatópontok felosztása három, illetve mindössze két csoport tételezése esetén.

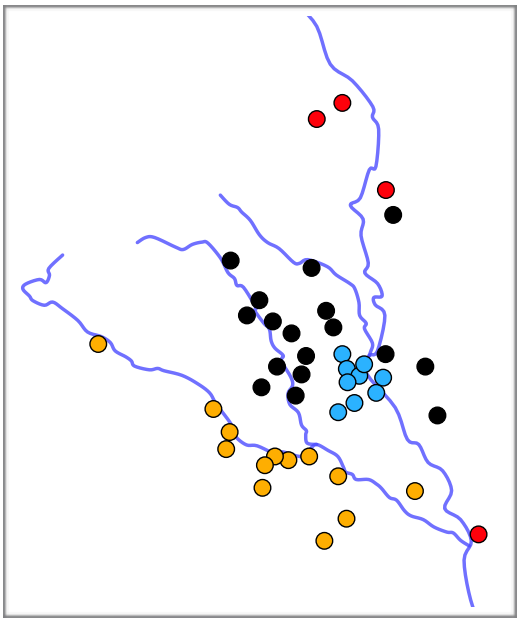
A klaszteranalízis eredményét – a kialakított csoportokat színekkel reprezentálva – térképre is vetíthetjük. Az 1.7. térkép a moldvai magyar nyelvjárások négy klaszteres felosztását szemlélteti, a MCsNyA. adataiból automatikus adatösszevetéssel készült mátrix alapján. A klaszteranalízis a Bihalbocsban generált hasonlósági mátrix alapján az R statisztikai programmal készült. A Bihalbocs képes fogadni és térképre vetíteni az elemzés eredményét, így – más megoldásokhoz, például egyéb dialektometriai szoftverek használatához képest – megmarad a koherencia a különböző térképes kimutatások ábrázolásmódjában.



1.9. ábra: A MCsNyA. finom fonetikai lejegyzése alapján készített hasonlósági mátrixon végzett Ward-féle klaszteranalízis dendrogramja

A klaszteranalízis eredményeképpen kialakított csoportok száma nem eleve adott, a kutató döntésén múlik, hány csoportot szeretne elkülöníteni. A moldvai nyelvjárások esetében az eddigi elemzések alapján, ahogyan az az 1.7. térképen is látható, négy terület elkülönítése látszik optimálisnak (Bodó et al. 2012, Bodó–Vargha 2016, vö. Péntek 2014). A négy csoport, földrajzi elhelyezkedése alapján, a következőképp nevezhető meg: északi (piros), Bákó környéki (kék), Tázló menti (fekete), Tatros menti (narancssárga). Dönt-hetünk azonban úgy is, hogy két, három vagy akár öt csoportot különítünk el és ábrázolunk térképen, ez a klaszteranalízis végeredményén valójában nem változtat, egy-szerűen csak annyit jelent, hogy az elemzés által kialakított csoportosítást (dendrogramot) milyen mélységig jelenítjük meg a térképen színekkel ábrázolva (a nyelvatla-szok adataiból készített klasztertérképekre számos példa található az 5. fejezetben).

A klaszteranalízisen alapuló nyelvjárási felosztások a hagyományos, nyelvjárásterületek, nyelvjárási régiók kijelölését célzó felosztásokat idézik, azzal a különbséggel, hogy itt nem a kutató által kiválasztott nyelvjárási térképek alapján meghúzott izoglosszák nyalábjainak segítségével jönnek létre a kijelölt területek, hanem egy adott korpusz valamennyi adata és azok valamennyi részletre kiterjedő tulajdonsága alapján. Vagyis végső soron maguk a lejegyzett adatok, illetve a bennük rejlő információ maga hozza létre a kijelölt területeket. A módszer előnye az objektivitás, hátránya, hogy a felosztás kategorikus, miközben a legújabb, magyar nyelvjárásterületeket megjelenítő térképek számolnak azzal is, hogy nem minden kutatópont sorolható be egyértelműen egyik vagy másik csoportba (l. Imre 1971: 333, Juhász 2001b: 460–461).



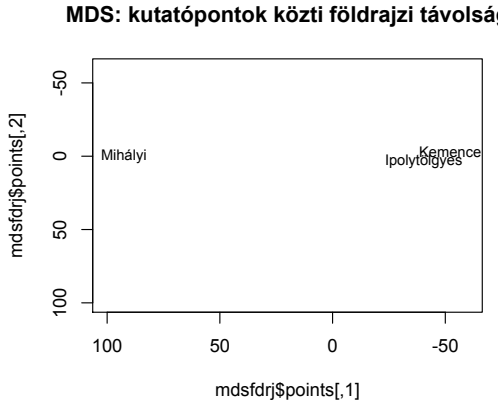
1.7. térkép: A moldvai nyelvjárások automatikus felosztása Ward-féle klaszteranalízissel a MCsNyA. adatainak dialektometriai elemzése alapján

A klaszteranalízissel létrehozott térképet jól kiegészítheti, ha többdimenziós skálázást (multidimensional scaling, MDS) is végzünk a mátrixunkon, illetve az elemzés eredményét színes térképen is szemléltetjük. A többdimenziós skálázás a klaszteranalízis kategorikusságával szemben a nyelvi tér kontinuum jellegét hangsúlyozza.

A nyelvi távolságot százalékértékekben mutató mátrixot nehéz lenne táblázatos formájában áttekinteni, különösen akkor, ha a kutatópontok száma (n) igen nagy. Az összes kutatópontpár közötti távolság (hasonlóság, eltérés mértéke) pontos ábrázolása egy n-1 dimenziójú „térben" történhet meg. Három kutatópont távolsági viszonyait tehát két dimenzió mentén (síkban, *x* és *y* koordinátaértékek kiszámításával), négy kutatópont esetében háromdimenziós térben (*x, y* és *z* értékek alapján) ábrázolhatjuk torzítás nélkül. Egy atlasz azonban nem három-négy kutatópontot tartalmaz: egy 99 kutatópontos atlasz (ilyen a S-ZA.) esetén 99-1, azaz 98 dimenziós „térre" volna szükség, ahol kutatópontonként a 98 koordináta (*x, y, z,* és így tovább) alapján történő ábrázolás gyakorlati szempontból teljesen megoldhatatlan. Mivel síkban vagy esetleg térben elhelyezkedő objektumok áttekintése lehet a célunk, szükségünk van egy olyan matematikai eljárásra, amely mátrixunk adatstruktúráját egy mindössze két- vagy háromdimenziós térbe transzponálja, gyakorlatilag is lehetővé téve az ábrázolást. Ez természetesen nem kivitelezhető bizonyos torzítás nélkül, ám az MDS olyan eszköz, amely két (*x, y*) vagy három (*x, y, z*) koordináta kutatópontokhoz való hozzárendelésével megfelel a nyelvész elvárásának: ami az n-1 dimenziós térben távol esik egymástól, az a redukált térben is – bizonyos hibahatáron belül – ugyanolyan távolságra lesz; a távolságok nagyságrendi viszonyai tehát megőrződnek.

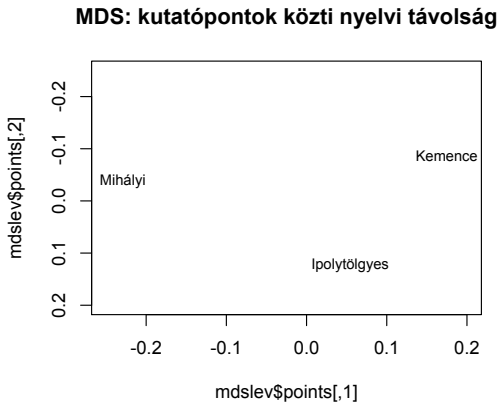
Az 1.10. ábra három MNyA. kutatópont, Mihályi, Ipolytölgyes és Kemence földrajzi távolságának kétdimenziós skálázását mutatja meg (hasonló egy szokásos földrajzi térképhez), az 1.11. ábra pedig ugyanezen kutatópontok nyelvi távolságáét (a három kutatópont földrajzi elhelyezkedését l. a 1.8. térképen). Jól látszik, hogy míg

Ipolytölgyes és Kemence földrajzilag közel helyezkednek el egymáshoz, addig a nyelvi távolság köztük jelentős, ami azzal magyarázható, hogy Kemence palóc település, Ipolytölgyes azonban nem. Mihályi földrajzilag mindkét kutatóponttól hasonló földrajzi távolságra van, nyelvileg azonban Ipolytölgyeshez közelebb áll, mint Kemencéhez. A kétdimenziós ábrázolás síkban mutatja meg, hogyan helyezkednek el egymáshoz képest egy nyelvjárási atlasz kutatópontjai a nyelvi térben. (A többdimeziós skálázás nyelvészeti felhasználásáról lásd Levshina 2016: 336–348 ; a dialektometriai mátrixok MDS-ének koordinátarendszerben történő ábrázolásáról lásd még Embleton et al. 2012.)

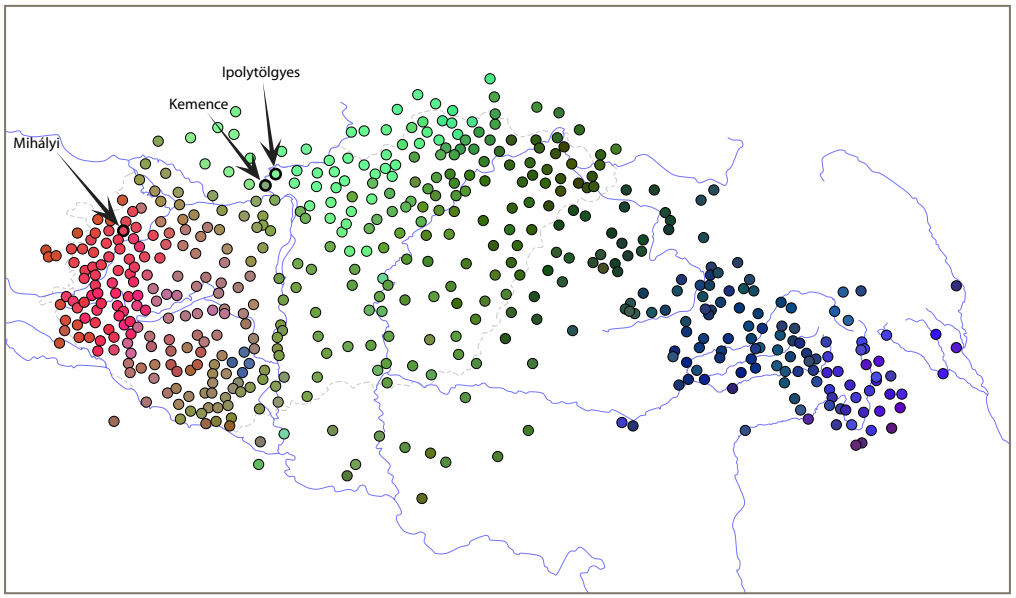


1.10. ábra: Három kutatópont földrajzi távolságának kétdimenziós skálázása

Három dimenzió már legalább az összvariancia 90%-át magyarázza dialektometriai mátrixok esetében, ahhoz azonban, hogy a kutatópontok közti összefüggéseket megjelenítsük, ebben az esetben három tengelyre, vagyis térbeli ábrázolásra van szükség. A településekhez kiszámított *x, y* és *z* értékeknek megfelelően azonban színkomponenseket is rendelhetünk a kutatópontokhoz, kiváltva ezzel a háromdimenziós euklideszi térben való megjelenítést; így a kutatópontok hagyományos térképes elhelyezkedése és a nyelvi hasonlóság összetett színekkel való ábrázolása egyszerre válik áttekinthetővé. Az egyes dimenziók így egy-egy színkomponensnek felelnek meg, amelyek intenzitása minden kutatópont esetében az adott dimenziónál meghatározott érték szerint alakul. A megjelenítéshez az RGB (red-green-blue) skálát használjuk, vagyis piros, zöld és kék színkomponenssel keverjük ki a kutatópontok összetett színárnyalatát (a módszerről l. Heeringa 2004: 156–164, 207–211, Leinonen 2010: 207-208). Az 1.8. térkép a MNyA. és a RMNyA. 482–482 integrált, finom fonetikai lejegyzésű térképlapjának elemzése alapján a teljes magyar nyelvterületet lefedő térképen mutatja be a háromdimenziós skálázás eredményét (a magyar nyelvjárási atlaszokról készített MDS-térképeket l. az 5. fejezetben).



1.11. ábra: Három kutatópont nyelvi távolságának kétdimenziós skálázása



1.8. térkép: A MNyA. és a RNyA. integrált dialektometriai elemzése többdimenziós skálázással készült térképen

Az 1.8. térképet a nagyobb magyar nyelvjárási régiók tükrében szemlélve legmarkánsabban a Nyugat-Dunántúl és a Palócföld mutat egyértelműen megragadható elkülönülést, a nyelvterület többi részét (a Partium és a Mezőség vagy a Mezőség és a Székelyföld között például) az átmenet jellemzi, a három komponensből kikevert színek nem látszanak élesen elkülönülni. A Székelyföld, főleg a keleti székely kutatópontok, lilás árnyalata a piros színkomponens határozott jelenlétére utal, számottevő nyelvi hasonlóságot sejtetve a Dunántúllal, ahol szintén hol markánsabban, hol kevésbé markánsan, de általában jellemző a pirosas színezet. A nyelvjárásszigetek színárnyalata a rokon nyelvjárások színével mutat hasonlóságot, legszembetűnőbben a

20. században áttelepített bukovinaiak és moldvaiak kutatópontjai a Dunántúl déli részén, amelyek kék színárnyalata leginkább a Székelyföld északkeleti kutatópontjaihoz hasonló, és nagyon határozottan elüt a közvetlen környezetüktől.

1.6. Összefoglalás

A fejezetben a dialektometriával kapcsolatos legalapvetőbb módszertani kérdéseket tekintettük át. Magát a dialektometriát a kvantitatív nyelvföldrajzi kutatási módszerek körébe soroltuk, lényege, hogy aggregált adatokkal dolgozik, vagyis több száz térképlap elemzésével hozza létre a kutatópontok közti nyelvi hasonlóságot vagy távolságot megmutató mátrixokat. Eszköztárát tekintve a számítógépes dialektológia része. Két alapvető módszert különböztettünk meg az adatok összevetésére, az egyik az előzetes, kutatói csoportosításon, munkatérképek létrehozásán alapuló salzburgi módszer, a másik a leginkább a groningeni egyetemhez köthető, az informatizált adatok automatikus összevetésén alapuló eljárás. A magyar nyelvjárási atlaszok informatizálása révén több százezer automatikusan elemezhető adat áll rendelkezésre, de a S–ZA., illetve a MNyA. és a RMNyA. integrált dialektometriai elemzése révén a két elemzési módszer összevetésére is lehetőség van. Megnéztük, hogyan lehet a nyelvi hasonlósági mátrixokat térképen megjeleníteni, a nyelvi hasonlóság mértékét színekkel érzékeltetve. Néhány példával szemléltettük, hogyan végezhetünk további statisztikai elemzéseket a kialakított mátrixokon (mátrixok közötti korreláció, a kutatópontok csoportosítására alkalmazható klaszteranalízis, a nyelvi kontinuumot megmutató többdimenziós skálázás).

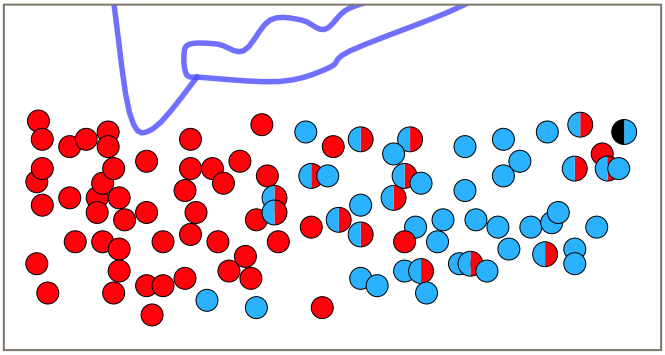
2. Elemzések abszolút sűrűségű kutatóponthálozaton – a Somogy-zalai nyelvatlasz[3]

A második fejezet a S–ZA. dialektometriai elemzését mutatja be. Az atlaszadatokból kétféle elemzés is készült, az egyik az adatok automatikus összevetésével, a másik a Király Lajos által készített csoportosítás alapján, az adattár formájában közölt térképlapokról. A két elemzés eredményét korrelációs elemzéssel és térképes elemzésekkel, korrelációs térképekkel vetjük össze. Fő kérdéseink: mennyire mutat hasonló képet a magyar egyezményes hangjelöléssel lejegyzett adatok automatikus adatösszevetése a kutatói csoportosításon alapuló elemzéssel, illetve az, hogy mennyiben térnek el egymástól a lejegyzés fonetikailag különböző mértékben részletes változataiból készített hasonlósági mátrixok?

2.1. Az adattár jellegzetességei, a térképlapok dialektometriai elemzése

Király Lajos atlasza gyűjtőterületéül a Balatontól délre, Somogy és Zala megye találkozását magában foglalóan jelölt ki egy nagyjából téglalap alakú területet. Tudatosan úgy alakította ki tehát a kutatóponthálózatot, hogy várhatóan jelentős nyelvjárási különbségeket tudjon feltérképezni a két megye között. Ezen fölül a jelenségek észak-déli megoszlására is eleve számított, leginkább az *ö*-zés vonatkozásában (Király 2005: 9–10.).

Az atlasznak összesen 99 kutatópontja van, a gyűjtés 1980 és 1985 között folyt. Az atlasz 204 címszó anyagát nem térképeken, hanem csoportosítva, szócikkek formájában tartalmazza. Az itt bemutatott dialektometriai elemzések ezekből a szócikkekből készültek, felhasználva Király Lajos kutatói csoportosítását. Az informatizált adattárban természetesen a nem térképes formában közreadott adatok éppen úgy kereshetők és térképezhetők, mint az adatbeírásos térképlapok adatai. A 2.1. térkép például a *csalán* hangtani változatait mutatja, a Király által meghatározott csoportosítási szempontok szerint, a mássalhangzók minősége alapján.

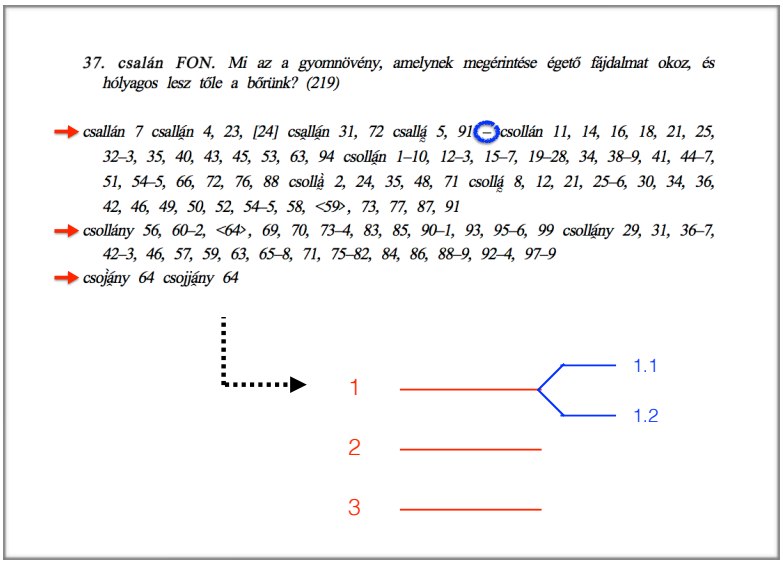


2.1. térkép: A *csalán* hangtani változatai Király csoportosítása szerint, piros = *csallán, csollán, csollá,* kék = *csollány,* fekete = *csojjány*

Az adattárszerű közlésmódhoz Király kétszintű csoportosítást alakít ki. Az adatokat külön bekezdésben fő csoportokra osztja, és amennyiben szükséges, kötőjelek segítségével alcsoportokat alakít ki a főcsoportokon belül. A *csalán* esetében például három

[3] A fejezet a Vékás Domokossal közösen tartott előadásunk (2009) és a 14. Methods in Dialectology konferencián elhangzott előadásom (2014) eredményeinek felhasználásával készült.

főcsoport van (ezeket jeleníti meg az 2.1. térkép), és az első (piros színnel jelölt) főkategórián belül van két alkategória, a magánhangzó minősége szerint külön csoportba kerülnek a *csallán* és a *csollán* adatok. Az adattárszerű közlést és csoportosítást a *csalán* címszó példáján az 2.1. ábra szemlélteti.



2.1. ábra: Az adatok csoportosítása a S–ZA. *csalán* szócikkében

A S–ZA. csoportosított térképlapjairól a fő- és alkategóriák alapján a salzburgihoz hasonló, kutatói csoportosításon alapuló módszerrel (ám munkatérképek létrehozása nélkül) készíthetünk dialektometriai elemzést. Az adatok egymáshoz képesti hasonlóságát így a fő- és alkategóriák alapján adjuk meg százalékos formában. Amennyiben két adat azonos főkategóriába és azonos alkategóriába tartozik (mint pl. az 1. ábrán szereplő *csollán* és a *csollà* változat), a köztük lévő hasonlóságot 100%-nak vesszük. Amennyiben a két adat azonos főkategóriába, de különböző alkategóriába tartozik (mint például a *csallán* és a *csollán* változat), a köztük lévő hasonlóság mértéke 50% lesz. A különböző főkategóriába tartozó adatok között (pl. *csollán* és *csollány*) a hasonlóság mértékét 0%-ban állapítjuk meg. Azon kutatópontok esetében, ahol több adat is előfordul, az automatikus adatelemzésnél is alkalmazott, az 1.4. fejezetben részletesen ismertetett eljárás szerint járunk el. A kutatópontok között az előforduló adatoknak megfelelően adjuk meg az átlagos hasonlóság mértékét, majd az összes térképlap alapján mért százalékos értékeket minden kutatópontpár esetében átlagoljuk, létrehozva a hasonlósági mátrixot (KL). Ily módon azon kutatópontpárok esetén lesz nagy a hasonlóság mértéke, ahol az adatok minél több esetben azonos fő- és alkategóriába esnek.

Ahhoz, hogy összehasonlítsuk a kutatói csoportosításon alapuló elemzés eredményét az adatok Levenshtein alapú összevetésével, a fonetikai részletesség különböző szintjeinek megfelelő mátrixokat alakítottunk ki. Azért többet, mert fontos annak eldöntése, milyen mértékű fonetikai pontosság szükséges az atlaszadatok dialektometriai felhasználhatóságához. Ehhez először is létrehoztuk az eredeti, finoman mellékjelezett lejegyzés graduálisan egyszerűsített formáit, majd az adatösszevetéseket ezeken is elvégeztük, az eredetivel együtt összesen négy hasonlósági mátrixot hozva létre.

A négy mátrix a fonetikai információ mennyiségének különböző szintjein mutatja meg a kutatópontok nyelvi hasonlósági viszonyrendszerét. Az első mátrix (LEV1) az eredeti, finoman mellékjelezett adatok alapján készült. A második esetben automatikusan töröltük az adatokról a mellékjeleket, a diftongusokat azonban megtartottuk, megszüntetve ugyanakkor a nyomatékeloszlásban lévő különbségek jelölését. A harmadik esetben további egyszerűsítéseket hajtottunk végre, összevonva egymással a hasonló magánhangzókat (az *ë*, *e*, *ä* és *é* hangokat, valamint az *à, a* és *á* hangot, illetve valamennyi hosszú magánhangzót annak rövid párjának feleltetve meg), a diftongusokat az azoknak megfelelő monoftongusokká (illetve azok rövid megfelelőjévé) alakítottuk át (így az *éi* diftongusból például *e* lesz). Legnagyobb mértékben a negyedik mátrix kialakításakor egyszerűsítettük a lejegyzést, ekkor valamennyi magánhangzót azonosnak vettük (*a*-nak feleltetve meg) és a hasonló képzésű mássalhangzókat is azonossá tettük (a zöngétlen mássalhangzókat azok zöngés párjával váltottuk föl, az *ly*-t *j*-nek feleltettük meg, az *r*-t *l*-nek, valamennyi nazálist *n*-nel helyettesítettük). A negyedik mátrix így már leginkább a lexikai szinten meglévő különbségekre érzékeny, a fonetikaiakra azonban már nem.

2.2. A dialektometriai elemzések eredménye, különböző mátrixok összevetése

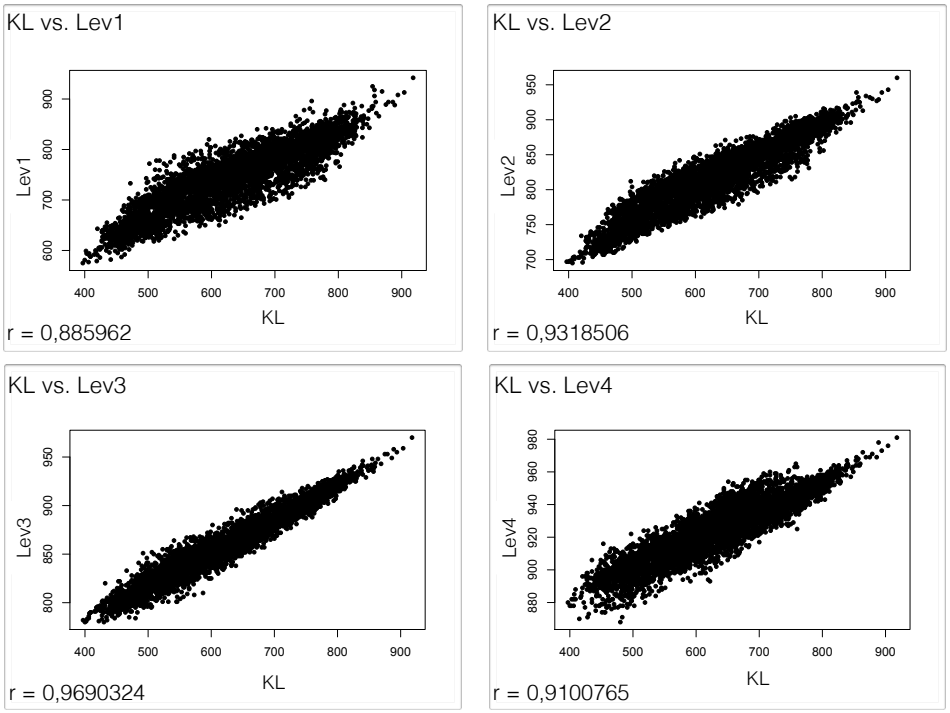
Az automatikus összevetéseken alapuló mátrixok a lejegyzés különböző finomságú szintjeinek esetében mért nyelvi hasonlóságot mutatják meg. A hasonlóság mértéke a kutatópontok között nem egyformán alakul a fonetikai információt különböző mértékben tartalmazó lejegyzések esetében. Alapvető különbség a négy hasonlósági mátrix között, hogy minél inkább egyszerűsítünk, annál nagyobb mértékben fognak hasonlítani egymásra az adataink, tehát annál nagyobb százalékos értékek szerepelnek majd a kutatópontpárok közti nyelvi hasonlóság mértékét mutató mátrixban. A 2.1. táblázat néhány példán keresztül megmutatja, hogyan alakulnak Büssü nyelvi hasonlósági viszonyai a Király-féle csoportosításból és a négy, automatikus adatösszevetések alapján készített mátrix szerint.

BÜSSÜ	KL	Lev1	Lev2	Lev3	Lev4
Kazsok	88,9	88,8	92,9	95,8	97,8
Mernye	80,2	76,6	85,6	91,5	94,7
Marcali	56,5	75,3	79,9	81,4	90,7
Alsórajk	40,0	58,1	69,7	78,0	87,9

2.1. táblázat: A nyelvi hasonlóság mértéke Büssü és négy kiválasztott kutatópont között különböző hasonlósági mátrixok szerint (százalékos értékek)

A fejezet legfontosabb kérdése, mennyire hasonlítanak egymásra a kutatói csoportosítás, illetve az automatikus adatelemzés alapján kirajzolódó nyelvi hasonlósági viszonyok. Ehhez először is a mátrixokat átfogóan összevető statisztikai elemzést célszerű végezni. Az egyik legkézenfekvőbb eljárás távolságmátrixok összevetésére a Mantel-teszt, amelynek segítségével nem csupán adatsorok, hanem mátrixok között végezhetünk korrelációs elemzést (Levshina 2015: 348–349). (A dialektometriai mátrixok esetében a távolság és a hasonlóság kiegészítik egymást, összegük mindig 100%, ezért az egyik a másikból egyszerűen kiszámítható, l. az 1.2. fejezetet).

A KL mátrixot a Lev mátrixokkal összevető Mantel-tesztekből az derül ki, hogy a kutatói csoportosítás eredménye leginkább a Lev3 mátrixszal korrelál, de a másik három mátrix esetében is igen erős az együttjárás, a korrelációs együttható értékei közelítenek az 1-hez, vagyis a teljes egyezéshez. Az összevetéseket a 2.2. ábra szemlélteti pontdiagramok segítségével, itt szerepelnek a Mantel-teszttel kiszámított korrelációs értékek is.



2.2. ábra: A Király csoportosítása alapján készült nyelvi távolságmátrix összevetése a Levenshtein-algoritmussal kiszámított távolság-mátrixszokkal

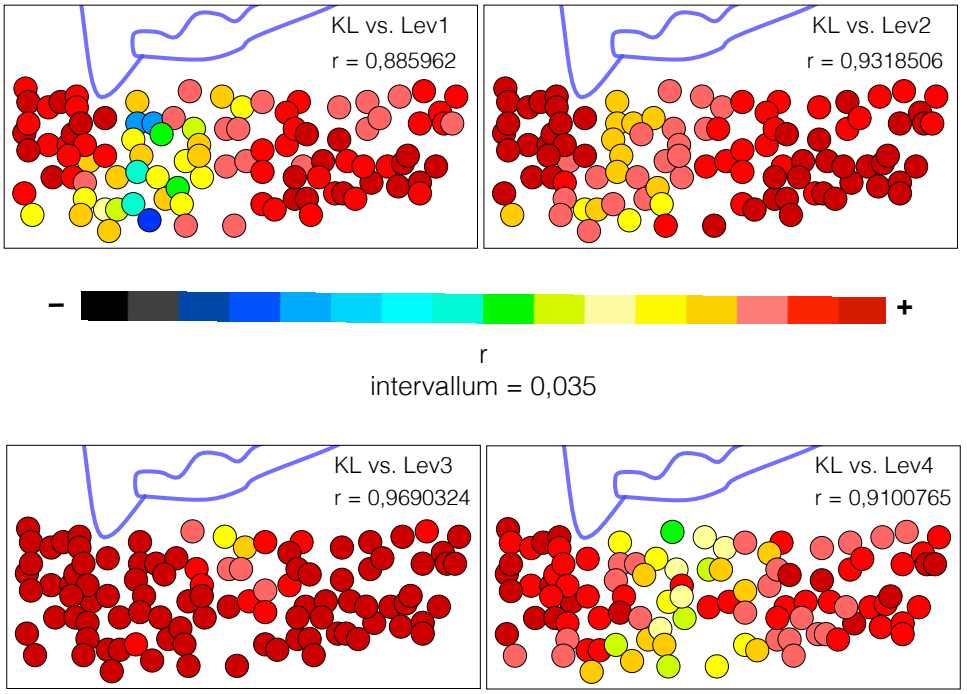
A korrelációs elemzés alapján elmondhatjuk, hogy az automatikus adatösszevetésen alapuló módszer a kutatói csoportosítást hasznosító módszerhez hasonlóan alkalmas a kutatópontok nyelvi hasonlóságának vizsgálatára. Megjegyzendő azonban, hogy az adatok egyszerűsítésével létrehozott mátrixok esetében – különösen a Lev2 és még inkább a Lev3 mátrixnál – valamivel nagyobb az együttjárás a csoportosítás alapján kiszámított KL mátrixszal. Feltehetőleg a jelenség azzal magyarázható, hogy a csoportosítás erősen sematizál, vagyis éppen az eredeti, finoman mellékjelezett lejegyzésre jellemző apró különbségek, amelyek számítanak a Lev1 mátrix kialakításakor, nem vesznek részt az adatok fő- és alcsoportokba való besorolásában, ahogy a 2.1. ábrán közölt példában is megfigyelhető.

Más módszerekkel vetnek össze kutatói csoportosításon alapuló és Levenshtein algoritmusával készített elemzéseket Valls és munkatársai (2012). Katalán dialektusokat vizsgálva nem találnak ilyen mértékű hasonlóságot a két elemzésmód adta eredmények között. Megjegyzendő azonban (egyéb, az elemzésre kiható tényezők mellett), hogy nem azonos korpuszon végzik a vizsgálatot. Feltehető, hogy azonos korpuszon vizsgálódva más atlaszok (esetleg más nyelvek) esetében is a S–ZA. atlaszon bemuta-

totthoz hasonló eredményekre juthatunk (a MNyA. és a RMNyA. kétféle módszerrel történő integrált elemzésére l. a 4.3. fejezetet).

A Mantel-teszt és a pontdiagramok tájékoztatnak arról, általában mennyire hasonlóan alakul a nyelvi távolság, a kutatópontok közti nyelvi hasonlósági viszonyrendszer különböző mátrixok esetében. Nem mutatja meg azonban azt, mely kutatópontoknál nagyobb az eltérés, illetve hogy melyek azok a kutatópontok, amelyeknél a választott módszertől függetlenül stabilak a hasonlóság földrajzi mintázatai.

Goebl különböző mátrixok összevetését olyan módon valósítja meg, hogy minden egyes kutatópont esetében megnézi a korrelációt, és aztán a korrelációs értékeket térképre vetíti (2005). A módszer lényege, hogy így nem csak a mátrixok között általában tudunk összevetést végezni, hanem egyenként vizsgálhatjuk a kutatópontok különböző dialektometriai elemzésekkel mért nyelvi hasonlósági viszonyai közti különbséget.

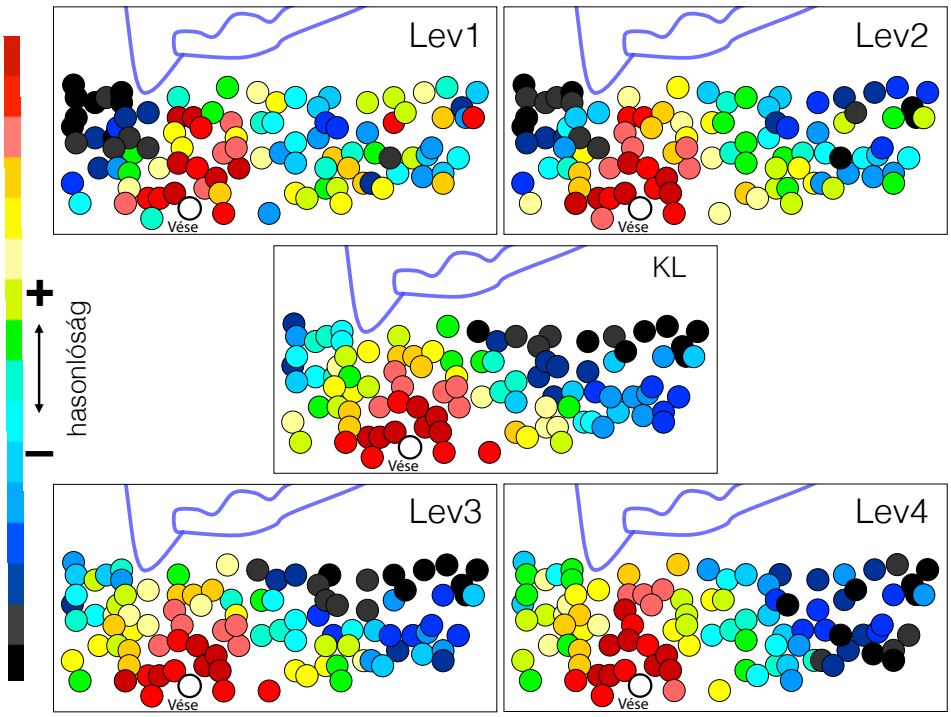


2.3. ábra: A Király csoportosításán alapuló (KL) és a Levenshtein algoritmussal készült mátrixokat összevető korrelációs térképek

Goebl korrelációs módszerét alkalmazva a 2.3. ábra a KL mátrix és a Lev mátrixok közötti Pearson-féle korrelációs együtthatók értékét kutatópontonként mutatja, térképen ábrázolva. A települések színe aszerint alakul, milyen mértékben korrelálnak az egyes kutatópontok esetében különböző módszerekkel mért nyelvi hasonlósági értékek (l. a színskálát). A Mantel-teszt eredményei és a pontdiagramok alapján nem meglepő, hogy a Lev3 mátrixszal való összevetéskor szinte csak piros kutatópontokat láthatunk, vagyis a KL és a Lev3 mátrix esetében kirajzolódó nyelvi hasonlósági mintázatok – két, északon, középen elhelyezkedő kutatópontot kivéve (Somogyszentpál és Táska) – szinte azonosnak mondhatók. Rendkívül erősek a korrelációk a KL és a

Lev2 mátrix között is, itt kissé kevésbé mutatnak egyezést a Somogy és Zala határában lévő kutatópontok. Általában a Lev1-es és a Lev4-es mátrixokkal való térképes összevetésen is a maximális egyezéshez közeli értékeket mutató piros színek dominálnak, a két megye határához közel azonban van néhány kutatópont, ahol számottevően kisebb a hasonlósági értékek közötti együttjárás. A legalacsonyabb a korrelációs együttható értéke Vése esetében (0,58), a KL és a Lev1 mátrix közti összevetéskor. Még két kutatópont van, ahol hasonlóan alacsony a korreláció mértéke: Somogysámson (r = 0,585) és Sávoly (r = 0,59). Az összes többi kutatópont esetében a korreláció 0,7 fölött van.

A korrelációs térkép megmutatja, mely kutatópont(ok) esetében találunk nagyobb eltérést, de azt már nem, pontosan hogyan, miben térnek el a nyelvi hasonlóság földrajzi mintázatai különböző dialektometriai elemzések esetében. Ahhoz, hogy erről képet kapjunk, külön kell térképeznünk az egyes kutatópontoknál kapott hasonlósági értékeket, különböző mátrixok használatával. Vése dialektometriai térképeit a 2.4. ábra szemlélteti.

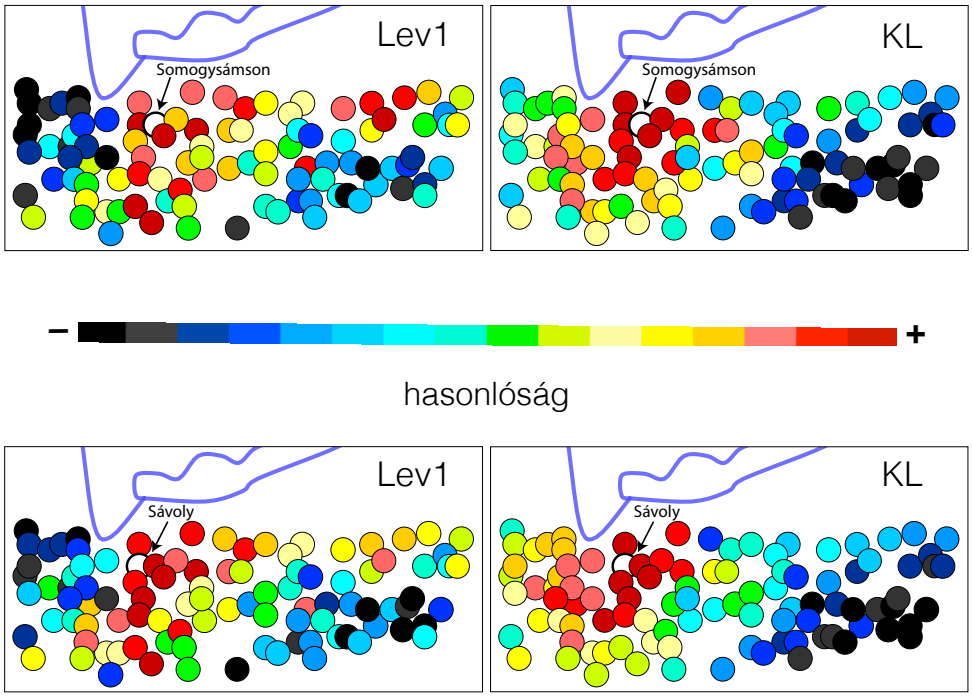


2.4. ábra: Vése dialektometriai térképei öt különböző mátrixszal nézve (a színek 6 kutatópontonként változnak, a színskála szerint)

Összevetve egymással a Vése dialektometriai térképein kirajzolódó földrajzi mintázatokat (2.4. ábra) megállapíthatjuk, hogy a KL és a Lev3 térkép szinte teljesen azonos, összhangban a két mátrixot összevető statisztikai térképpel (2.3. ábra), illetve nem látszik jelentős különbség a KL és a Lev4 térkép között sem. Mindhárom esetben a Véséhez földrajzilag közeli kutatópontok mutatnak a leginkább hasonlóságot, és leginkább az atlasz gyűjtőterületének keleti részén lévő kutatópontok nyelvjárása tér el. A Lev2 térképen már nem olyan nagy az eltérés a keleti kutatópontok és Vése között, a zalai kutatópontok esetében viszont csökken a nyelvi hasonlóság. Leginkább a Lev1

térkép különbözik a KL térképtől, szintén összhangban a korrelációs elemzés eredményével. A különbség elsősorban nem a közeli, inkább a távolabbi kutatópontokat érinti. A Lev1 térkép szerint Vése nyelvi hasonlósági kapcsolatai inkább keleti irányba, vagyis Somogy felé mutatnak, míg a KL térképen a földrajzilag közelebbi, zalai pontok hasonlítanak jobban.

A Lev1 mátrix abban különbözik az összes többitől, hogy a hátterében lévő elemzéskor az eredeti, finom fonetikai részleteket is tartalmazó lejegyzésben vetjük össze az adatokat, vagyis a mellékjelek is részt vesznek az összevetésekben. Az a földrajzi mintázat tehát, amit látunk, az apró hangtani részletek figyelembevételén alapul. A 2.4. ábrán szereplő dialektometriai térképek összevetése alapján csak ezen az elemzési szinten mutatkozik nagyobb hasonlóság Vése és az atlasz keleti kutatópontjai között.

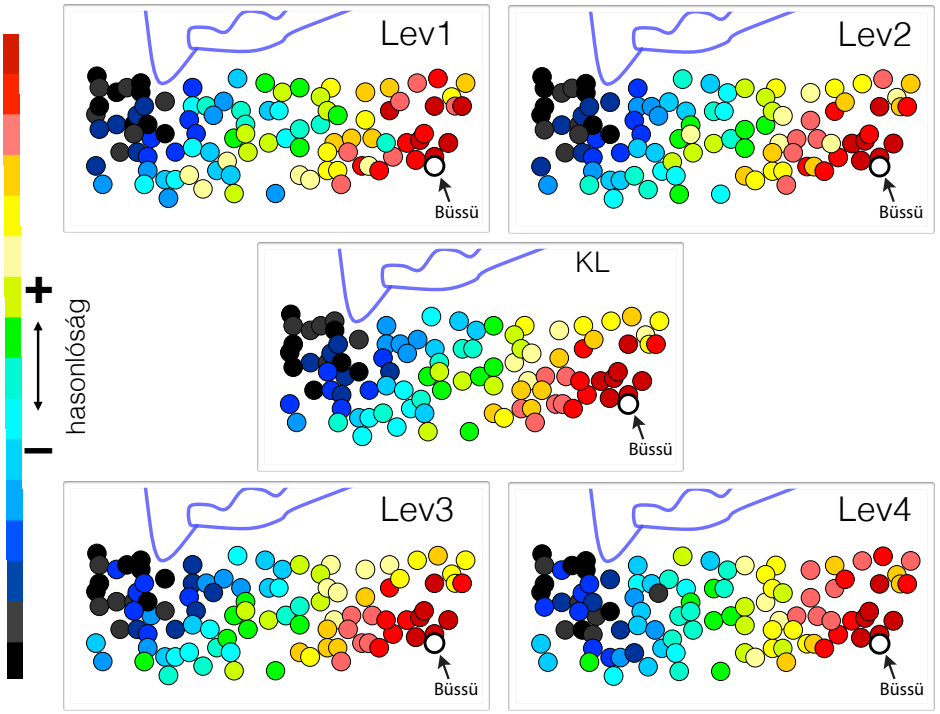


2.5. ábra: Somogysámson és Sávoly Lev1 és KL mátrix alapján készített dialektometriai térképei (a színek 6 kutatópontonként változnak, a színskála szerint)

Somogysámson és Sávoly a két másik kutatópontunk, ahol a Lev1 és a KL mátrix közötti korreláció számottevően kisebb, mint más települések esetében. Összevetve egymással Lev1 és KL mátrix alapján készült dialektometriai térképeiket azt látjuk, hogy a Vése esetében tapasztaltakhoz hasonlóan a nyugati (zalai) és a keleti (somogyi) kutatópontok között mutatkozik „váltás”. A különbség különösen jelentős Sávoly esetében, ahol a kutatóponttól nyugatra lévő, a Lev1 térképen kék és fekete színekben látszó kutatópontok a KL térképen pirosas árnyalatokat vesznek föl, vagyis a kisebb nyelvi hasonlóságot mutató kutatópontok csoportjából átkerülnek a nagyobb hasonlóságot mutató kutatópontok csoportjába (l. a 2.5. ábrát)

Összevetésül érdemes megnéznünk legalább egy olyan kutatópontot is, ahol a korrelációs elemzés alapján nincs számottevő különbség a KL mátrix és a Lev mátrixok között. Ilyen kutatópont például a vizsgált terület délkeleti sarkában elhelyezkedő

Büssü. Ha összevetjük egymással Büssü dialektometriai térképeit (2.5. ábra), azt látjuk, hogy valóban rendkívül hasonlóak, a Lev1 mátrix sem különbözik számottevően a többitől. Mind a négy térképen a földrajzilag közeli, somogyi kutatópontok mutatnak a leginkább hasonlóságot.



2.6. ábra: Büssü dialektometriai térképei különböző mátrixokkal nézve (a színek 6 kutatópontonként változnak, a színskála szerint)

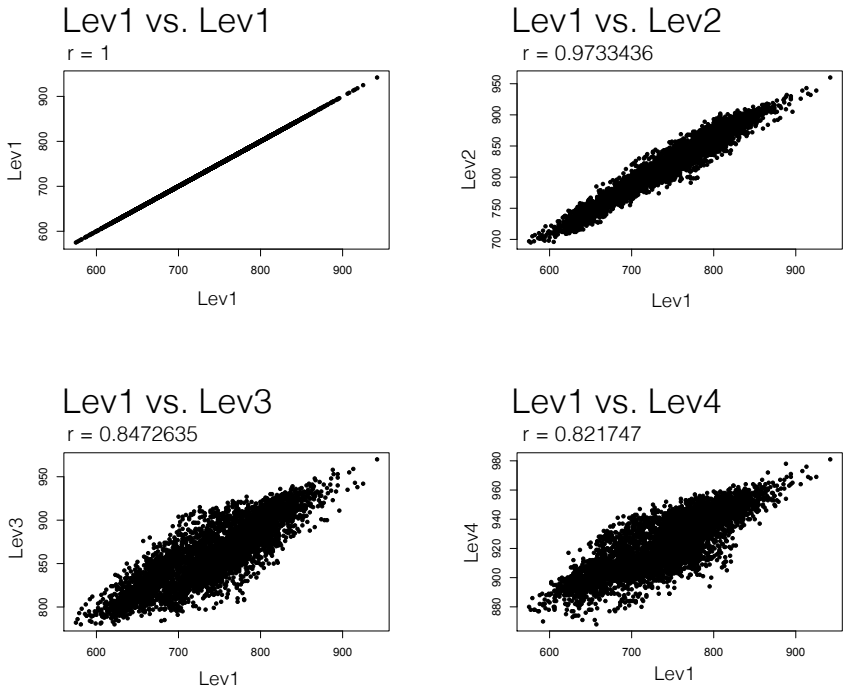
A korrelációs térképek (2.3. ábra) alapján feltételezhetjük, hogy a S–ZA. kutatópontjainak Levenshtein módszerrel készített dialektometriai elemzései általában, Büssühöz hasonlósan, nem mutatnak számottevő különbséget a Király-féle elemzéshez képest, és ez még a finom fonetikai különbségeket figyelembe vevő (Lev1) elemzésre is igaz. A Lev1 és a KL mátrix között Mantel-teszttel mért, a többinél kisebb korrelációs érték feltehetően néhány, a megyehatárhoz közeli kutatópont nyelvi hasonlósági viszonyaiban lévő különbségnek tudható be. Összességében elmondhatjuk, hogy mindkét elemzési módszer, a csoportosítás alapú és az automatikus fonetikai elemzés is, hasonló módon alkalmas a kutatópontok közti nyelvi hasonlósági viszonyok feltárására.

2.3. Fonetikailag érzékeny és lexikai jellegű mátrix összevetése

A KL mátrixszal való összevetéshez összesen négy, fonetikai különbségekre különböző mértékben érzékeny, Levenshtein alapú mátrixot hoztunk létre (l. a 2.1. fejezetet). A mátrixok közül az első (Lev1) az apró fonetikai különbségek figyelembevételével jött létre, ez tekinthető a leginkább fonetikai alapúnak. A negyedik mátrix kialakításakor (Lev4) oly mértékben egyszerűsítettük az eredeti lejegyzést, hogy az már lényegében a lexikai különbségek nyelvi hasonlósági mintázatokra gyakorolt hatását tükrözi. A következőkben azt vizsgáljuk meg közelebbről, mennyiben hasonlít, illetve mennyiben különbözik egymástól a fonetikailag érzékeny (Lev1) és a lexikai

jellegű (Lev4) elemzés, illetve a nyelvi hasonlósági viszonyok mennyiben alakulnak másképpen a lexikai, mint a fonetikai jellegű mátrix alapján.

Elsőként a Lev1 mátrixot célszerű összevetnünk Mantel-teszt és pontdiagramok segítségével a Lev2, Lev3 és Lev4 mátrixszal, hasonlóképpen a többi mátrixot a KL-lel összevető elemzéshez (a KL mátrix korrelációs elemzéseit l. 2.2. ábrán). A Lev1 mátrixot saját magával és a többi Levenshtein alapú mátrixszal összevető pontdiagramokat és a Mantel-teszt eredményeképpen kapott korrelációs értékeket a 2.7. ábra szemlélteti.

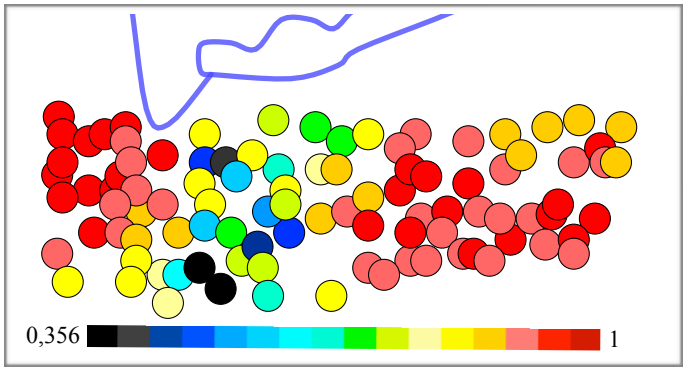


2.7. ábra: A Lev1 és a többi Levenshtein algoritmus használatával készült mátrix korrelációja

A Lev1 mátrix saját magával értelemszerűen azonos (az r = 1 érték maximális hasonlóságot jelez). Az első pontdiagram mindössze annyi információval szolgálhat, hogy megnézhetjük rajta, milyen mértékben egyeznek az adatok a legnagyobb, illetve a legkisebb hasonlóságot mutató kutatópontok esetében. A Lev1 és a Lev2 mátrix összevetése alapján a finom fonetikai és a mellékjeleket nem tartalmazó elemzés szinte azonos eredményt ad. Nagyobb különbség van a Lev1 és a Lev3 mátrix között, és valamivel ennél is jobban különbözik egymástól a Lev1 és a Lev4 mátrix, vagyis a fonetikai és a lexikai jellegű mátrix között találjuk a legnagyobb különbséget (noha még ezekben az esetekben is erős korrelációról beszélhetünk az r értéke alapján).

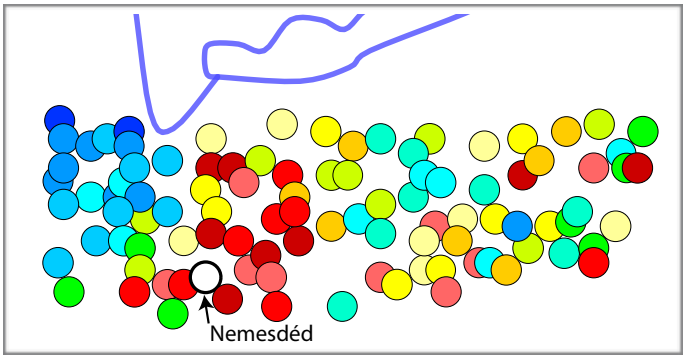
Azt, hogy pontosan mely kutatópontok esetében találunk nagyobb különbséget a fonetikai és a lexikai jellegű összevetés között, korrelációs térképen nézhetjük meg. A Lev1 és Lev4 mátrix közti kutatópontonkénti korrelációs értékeket a 2. térkép szemlélteti. A legalacsonyabb értékeket Nemesdédnél (r = 0,356) és Vésénél (r = 0,422) találjuk, de kisebb mértékű a két mátrix közti együttjárás Somogysámson (r = 0,478) és Sávoly (r = 0,579) esetében is, csakúgy, mint a KL és a Lev1 mátrix összevetésekor. Szintén alacsonyabb a korreláció további két kutatópont, Szenyér (r = 0,513) és

Mesztegnyő (r = 0,563) esetében. A gyengébb korrelációt mutató kutatópontokban közös, hogy földrajzilag nagyjából középtájt helyezkednek el, vagyis a megyehatárhoz közelebbi somogyi településekről van szó.



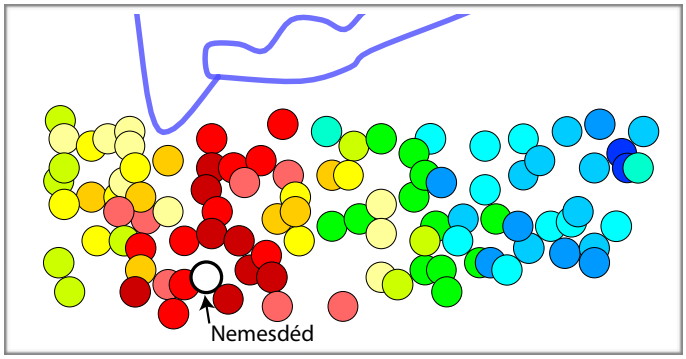
2.2. térkép: Kutatópontonkénti korreláció a Lev1 és a Lev4 mátrix között. A kutatópontok színe a korreláció mértéke szerint alakul, egy szín 0,035-nyi intervallumnak felel meg, a vöröstől a feketéig, a színskála alapján

Ahhoz, hogy lássuk, pontosan milyen különbségek vannak az egyes kutatópontok nyelvi hasonlósági viszonyaiban, el kell készítenünk a konkrét kutatópontok dialektometriai térképeit. A korrelációs együttható alapján azt várhatjuk, hogy Nemesdéd fonetikai és lexikai jellegű hasonlósági térképe kevéssé hasonlít egymásra (2.3. és 2.4. térkép).

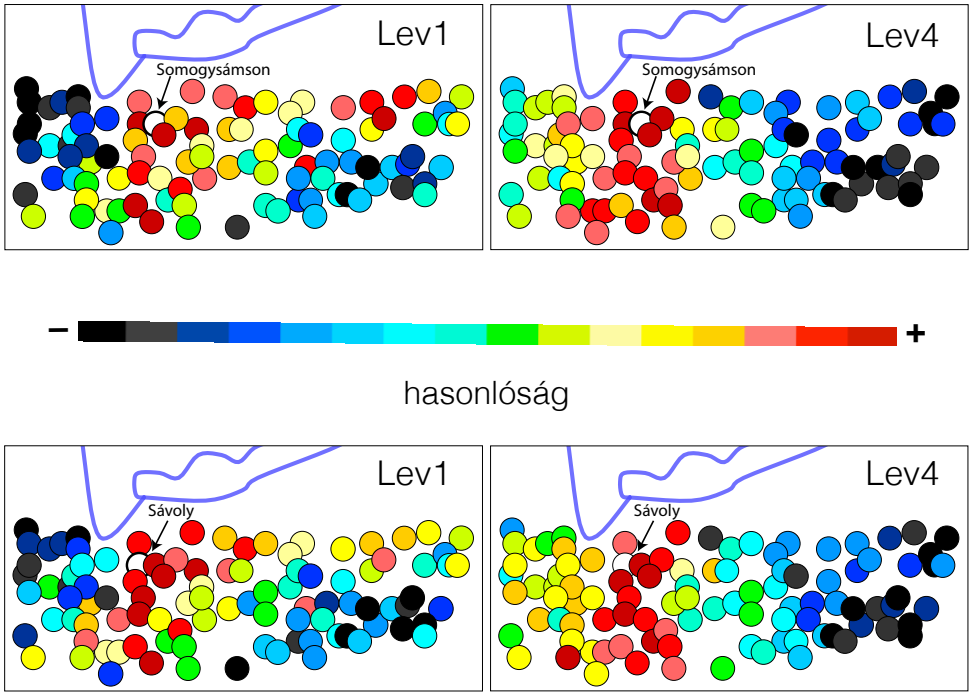


2.3. térkép: Nemesdéd dialektometriai térképe a fonetikailag érzékeny (Lev1) mátrix alapján (a színek 8 kutatópontonként változnak)

Nemesdéd különböző mátrixok alapján készített térképei valóban jelentős eltérést mutatnak abban a tekintetben, hogy nyugati vagy keleti irányba mutat-e inkább a nyelvi hasonlósági súlypontja. Hangtanilag inkább keleti, lexikailag azonban inkább nyugati jellegűnek mutatkozik. A vele leginkább hasonlóságot mutató kutatópont azonban nem változik meg, ez mindkét esetben Vése, de vannak tőle távolabbi kutatópontok is, szintén a térkép középső sávjában, amelyek mindkét mátrix alapján nagyobb hasonlóságot mutatnak, Somogysámson és Sávoly.



2.4. térkép: Nemesdéd dialektometriai térképe a lexikai jellegű (Lev4) mátrix alapján (a színek 8 kutatópontonként változnak)



2.8. ábra: Somogysámson és Sávoly dialektometriai térképei a fonetikai (Lev1) és a lexikai (Lev4) mátrix alapján (a színek 6 kutatópontonként változnak)

Vése különböző mátrixok alapján készített térképeit a 2.4. ábra szemlélteti. Összevetve a Lev1 és a Lev4 térképet, Vése nyelvi hasonlósági viszonyai rímelnek a Nemesdéd esetében látottakra: a Lev1 térképen látszó keleti hasonlóságok eltűnnek, a nyugatiak fölerősödnek. A 2.8. ábra a Lev1 és Lev4 térképeket szemlélteti Somogysámson és Sávoly esetében. A nyelvi hasonlóság földrajzi mintázatai itt is éppen úgy alakulnak, mint a fönti, Nemesdéd és Vése nyelvi hasonlósági viszonyrendszereit ábrázoló térképeken.

Vése, Sávoly és Somogysámson, de még Nemesdéd is – a korrelációs értékek alapján – a nagyobb eltérést mutató kutatópontok közé tartozik. Korábban láttuk, hogy éppen ezen kutatópontoknál nagyobb eltérés mutatkozik a Lev1 és a KL mátrix között is, ugyanakkor a Lev4 mátrixszal való összevetésben csak Nemesdéd mutat valamivel jelentősebb különbséget.

Az összevetések alapján logikus magyarázatnak tűnik, hogy vannak olyan somogyi kutatópontok, amelyek hangtanilag őrzik somogyi jellegüket, de lexikailag már inkább zalaiak, vagy legalábbis igen erős zalai hatás tapasztalható náluk a lexikont illetően. A KL mátrixszal való korrelációs elemzések arra engednek következtetni, hogy Király csoportosításában, mivel (ahogyan ezt föntebb is megállapítottuk) kevésbé érvényesül a finom fonetikai jellegzetességek hatása, a kutatópontok közti hasonlóság megállapításakor a lexikai hasonlóság nagyobb súllyal esik a latba.

2.4. Összefoglalás

Két dialektometriai módszert, az adatok csoportosításán és a Levenshtein-algoritmussal történő automatikus összevetésén alapulót, hasonlítottunk össze korrelációs elemzés és térképek segítségével. A Levenshtein alapú elemzéshez négy, a lejegyzés fonetikai tartalmának különböző szintjeit reprezentáló mátrixot is készítettünk. Az eredmények alapján mindkét módszer egyaránt alkalmas a kutatópontok közti nyelvi hasonlóság vizsgálatára. Van azonban néhány kutatópont, a megyehatár közelében, amelyek jelentősebb eltérést mutatnak, főképp a finoman mellékjelezett lejegyzésen (Lev1) és a csoportosításon (KL) alapuló mátrix között. Ezen kutatópontoknál nagyobb eltérés figyelhető meg akkor is, ha a fonetikailag érzékeny mátrixot a lexikai jellegűvel vetjük össze. A megyehatárhoz közel tehát számos olyan település található, amelyeknek ellentétes irányba mutatnak a fonetikai és a lexikai hasonlósági viszonyai. A különböző mátrixok korrelációs elemzéseinek eredményei arra engednek következtetni, hogy a Király-féle kutatói csoportosítás esetében az árnyalatnyi fonetikai különbségek kevésbé játszanak szerepet az aggregált hasonlósági értékek alakításában, erőteljesebben érvényesül a lexikai változatok közti eltérések hatása.

3. Fonetikai információ, településtörténet, földrajzi távolság – A magyar nyelvjárások atlasza[4]

Ebben a fejezetben azt a kérdést járjuk körül a MNyA. informatizált változatának többszintű dialektometriai elemzésén keresztül, hogy a különböző nyelvi szintek (lexikon, hangtan, illetve a különböző elvontsági szinten megragadott hangtani jelenségek) a folyamatos diakrón változás közepette mennyire stabilak, abban az értelemben, hogy mennyire állnak ellen a környező nyelvváltozatok hatásának. Feltételezhetjük, hogy a különböző nyelvi szintek eltérően reagálnak a környező nyelvjárások hatására, azzal összefüggésben, hogy elemeik milyen mértékben állnak egymással kölcsönhatásban, tehát mennyire alkotnak szoros, illetve laza rendszert (az előbbit a hangtan, utóbbit a lexikon példázza). E kérdést leginkább olyan beszélőközösségek esetén lehet módszeresen vizsgálni, amelyek az eredeti nyelvjárási környezetükből kiszakadva egy másik, a korábbitól jellegzetesen eltérő nyelvjárási környezetbe kerültek, ahol rövidebb-hosszabb ideig kontaktushatásoknak voltak kitéve.

3.1. Az adattár jellegzetességei a dialektometriai elemzés szempontjából

A MNyA. előkészítésekor, az adatgyűjtés szakaszában két kérdőívvel dolgozott a munkacsoport. Csak az első, többségében hangtani szempontból érdekesnek gondolt kérdőívet kérdezték ki az atlasz valamennyi kutatópontján, a második, elsősorban lexikai változókra koncentráló kérdőívet már csak a kutatópontok egy részén (Deme 1975). A dialektometriai elemzéshez elengedhetetlen, hogy az összevetéseket az adatmennyiség szempontjából kiegyenlített térképlapokon végezzük, hiszen ha nem így járunk el, tulajdonképpen több alkorpuszon végzünk összevetéseket, és a hasonlósági mátrixunk nem lesz koherens, vagyis nem lesz alkalmas arra, hogy általa a különböző kutatópontok nyelvi hasonlósági viszonyai alapján stabilnak látszó földrajzi mintázatokat rajzoljunk meg (vö. Vargha 2013). Így tehát csak azok a térképlapok kerülhettek bele az elemzett korpuszba, amelyeket az atlasz valamennyi (N = 395) kutatópontján fölgyűjtöttek, és a teljes gyűjtőterületen vizsgálható jelenségeket mutatnak be (vagyis elenyésző adathiányt és kevés bizonytalan hitelességű adatot tartalmaznak).

A dialektometriai kutatásokban 100 térképlapnyi adatmennyiséget szokás minimumnak tekinteni az esetlegességek hatásának minimalizálása érdekében, hogy néhány (pl. lejegyzési) anomália ne befolyásolhassa érdemben az eredményeket, és hogy elemzéseink alapján valóban releváns összefüggéseket tárjunk föl (vö. Heeringa 2004). Jelen kutatásban ennél jóval több, az atlasz összesen 1162 térképlapja közül 645 felel meg a fenti kritériumoknak. Az elemzések ennek megfelelően megbízhatónak tekinthetők a magyar nyelvjárások közti nyelvi hasonlósági viszonyokat illetően, az atlasz kutatópont-hálózatának viszonylatában.

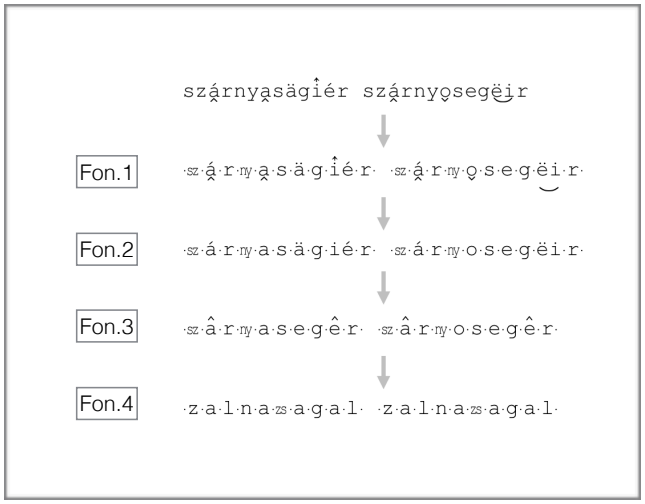
3.2. Az elemzési módszerről

A lejegyzésben rejlő információgazdagság hatásának vizsgálatához, valamint a fonetikai, fonológiai és lexikai szinten megmutatkozó nyelvi hasonlósági viszonyok összevetéséhez az eredeti lejegyzésen túl automatizált eljárással létrehoztam további három, az eredetinél információszegényebb lejegyzési formát. A lejegyzés különböző formái közti különbségeket az 1. ábra szemlélteti, ahol a felső sorban két, informatizált atlaszadat szerepel, az adatrögzítés során alkalmazott, szerkeszthető formában.

---

[4] A fejezet a Nyelvelmélet és dialektológiai 3. kötetben megjelent tanulmányom (Vargha 2015a) bővített, átdolgozott változata.

A Fon.1 szint az első elemzési szint, ebben az esetben nem változik a lejegyzésben rejlő információ mennyisége, a kidolgozott algoritmus csupán összevonja az azonos funkciójú, csak grafikusan különböző mellékjeleket, továbbá egyeleműve teszi a szerkeszthető változatban többelemű mássalhangzókat (pl. *sz, cs, dzs*), így biztosítva, hogy az elemzés során valóban hangokat vessünk össze egymással, és ne grafikus elemeket.



3.1. ábra. Példa a lejegyzés információgazdagságára különböző elemzési szinteken

A Fon.2 szinten automatikusan eltávolítjuk a mellékjeleket, de meghagyjuk a diftongusokat és valamennyi különböző alapjelet. Míg a Fon.1 a finom fonetikai szint, a Fon.2 durvább, de még mindig jellegzetesen fonetikainak minősülő szint.

A Fon.3 lejegyzési szint a „fonológiai" szint. Célom az egyszerűsítéssel az, hogy egy lehetséges fonológiai jellegű lejegyzést jól közelítő formát hozzak létre, amelyből hiányoznak a részletes fonetikai információk. Az adatok automatikus egyszerűsítése fonológiai jellegű lejegyzéssé több okból sem lehet tökéletes. Az egyszerűsítő algoritmus egyaránt *é*-vé alakítja az egymással néhány észak-keleti kutatóponton oppozícióban lévő, a köznyelvi *é* helyén álló középső nyelvállású palatális illabiális diftongusokat, pl. *széil* 'légáramlat' : *sziél* 'szegély' (az észak-keleti nyelvjárások különböző *é* fonémáiról lásd bővebben N. Fodor 2012). Ugyanez történik valamennyi *ou* és *uo*, valamint *öü* és *üö* diftongussal is, azok területi elhelyezkedésétől függetlenül. A félhosszú hangok nem feleltethetők meg egyértelműen sem rövid, sem hosszú hangoknak, ezért félhosszúak maradnak, hiszen bármilyen irányú besorolásuk önkényes volna. Megemlítendő továbbá, hogy a polifonémikus, hosszú magánhangzóként megjelenő hangokra az automatikus módszer nincs tekintettel. Így például az *āra* és az *ára*, illetve a *mēj* 'mell' és a *méj* 'mély' egybeesik. Ez azonban feltehetőleg egyáltalán nem, vagy csak igen kivételes esetekben járhat olyan eredménnyel, hogy az elemzés során valóban azonosnak vennénk különböző fonémákat, hiszen egyszerre csak egy térkép adatait, tehát egy nyelvi jelenség különböző változatait vetjük egymással össze.

Egyebekben a nyelvjárási hangok fonémaszerű elemeknek való megfeleltetése a Magyar dialektológiában leírtaknak megfelelően történik (Fodor 2001: 327–334): a palóc nyelvjárásokban az illabiális *ȧ* és a labiális hosszú *ā* fonémaértéke nem különbözik más nyelvjárások *a* és *á* hangjaiétól; a nyílt *ä* nem különbözik az *e* hangtól, a zárt *ë* viszont igen; a diftongusokat hosszú monoftongusoknak feleltetjük meg. Összességében elmondható, hogy a kidolgozott algoritmus eredménye jól közelíti a magyar

dialektológiában hagyományosnak tekinthető fonológiai lejegyzést, ugyanis a hiányosságok alacsony száma kvantitatív keretben elfogadható.

Az itt alkalmazott egyszerűsítési módszer nem jelent konkrét kötődést egyetlen fonológiai irányzathoz sem; annyiban nevezhető általában véve fonológiai jellegűnek, amennyiben a legtöbb fonológiai megközelítés sajátja, hogy minél több fonetikai részletet redundánsnak, irrelevánsnak tekint, amelyektől elvonatkoztat.

A Fon.4 szintű egyszerűsítés során valamennyi magánhangzót azonosnak vesszük, továbbá a hasonló mássalhangzókat is egybevonjuk (pl. zöngés-zöngétlen különbség megszüntetése; *l* és *r* azonos lesz), így tulajdonképpen drasztikusan csökkentjük a kiejtésben mutatkozó különbségek súlyát, és viszonylagosan fölerősítjük a lexikai azonosságok és eltérések hatását az elemzésben.

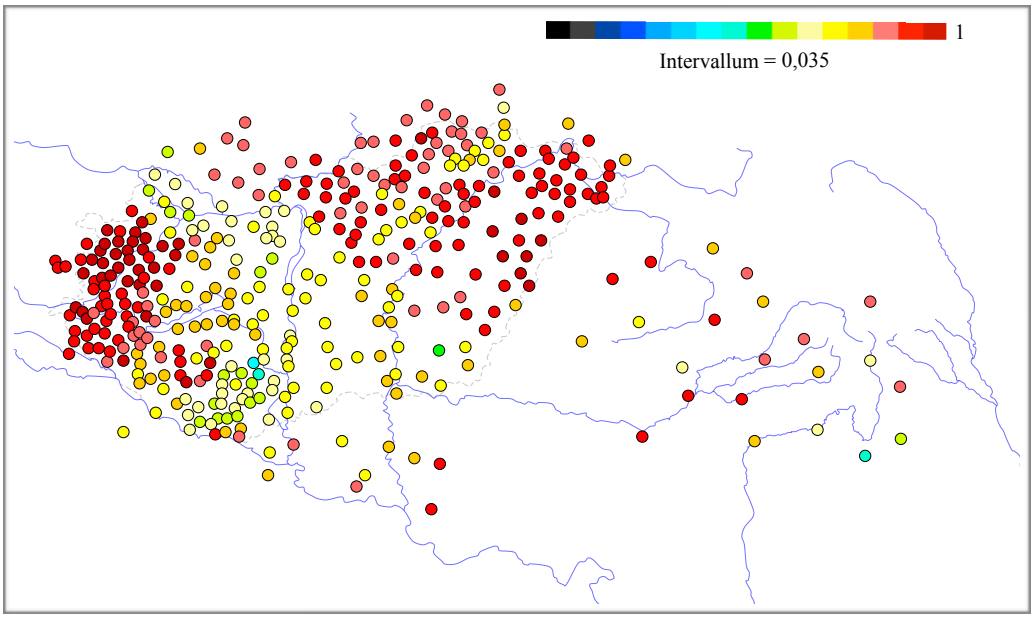
A hangminőségekre vonatkozó információkat különböző mértékben tartalmazó négy szint mindegyikén külön elvégezve az elemzést négy hasonlósági mátrixot kapunk. A mátrixokat Mantel-teszt és korrelációs térképek segítségével vetjük össze, valamint az elemzés szempontjából legérdekesebb kutatópontok esetében az egyes kutatópontok hasonlósági viszonyait megmutató térképeket is összehasonlítjuk, így fényt deríthetünk arra, hogy azok a nyelvi szintek – fonetikai (Fon.1, Fon.2), fonológiai (Fon.3), lexikai (Fon.4) –, amelyeket a különböző lejegyzési szintekhez, illetve mátrixokhoz társítottunk, mennyire alkalmasak az adott kutatópont nyelvi hasonlósági viszonyrendszerének feltárására.

3.3. Mátrixok közti korreláció

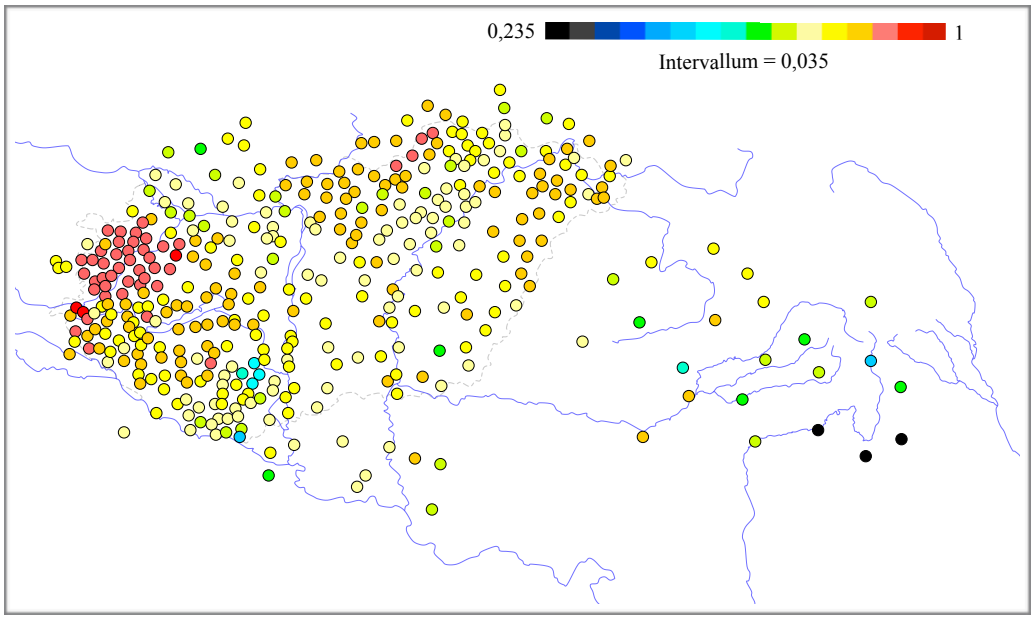
Elsőként a három módosított lejegyzés alapján készült hasonlósági mátrixot (Fon. 2, Fon.3, Fon.4) érdemes Mantel-teszt segítségével összevetnünk az eredeti, valamennyi fonetikai részletet tartalmazó lejegyzés alapján készült mátrixszal (Fon.1). A többi mátrix korrelációja a Fon.1 mátrixszal azok sorrendjét követve a következőképpen alakul: 0,874, 0,820, 0,790. Általánosságban azt állapíthatjuk meg, hogy a nyelvi hasonlósági viszonyokra nem feltétlenül van hatással, milyen adatokon végezzük az elemzést: a leginkább csak a fonológiai (Fon.3), illetve lexikai (Fon.4) különbségekre érzékeny változatok alapján létrehozott mátrixok is erősen korrelálnak az eredeti, mellékjeleket is tartalmazó lejegyzésből készült nyelvi hasonlósági mátrixszal (Fon.1).

Ahhoz, hogy megtudhassuk, mit jelent ez az egyes kutatópontokra nézve, az előző fejezetben is alkalmazott módszerrel a mátrixok közötti összevetéseket vizualizáló korrelációs térképeket készítettem. A módszer lényege, hogy valamennyi kutatópont esetében összevetjük a kétféle mátrixban megadott hasonlósági értékeket, kiszámítjuk a Pearson-féle korrelációs együtthatót. Az így kiszámított értékeket egy színskálának megfelelően térképezzük. A térkép megmutatja, hogy a különböző mátrixok hátterében álló elemzésmódok által kialakított nyelvi hasonlósági viszonyrendszerek milyen mértékben mutatnak együttjárást az egyes kutatópontokra nézve (a módszerről részletesen l. Goebl 2005).

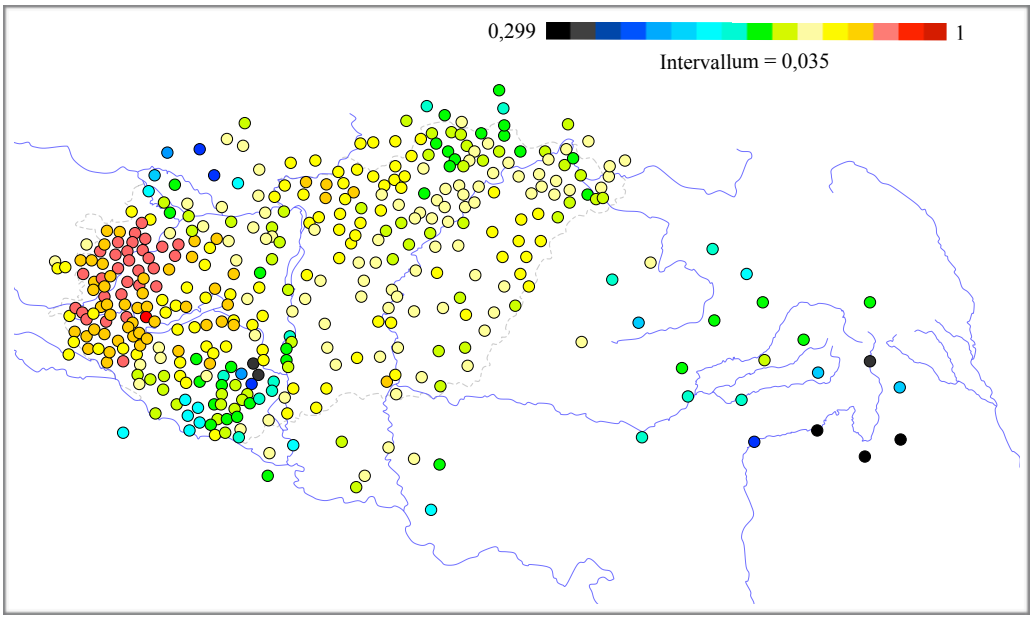
A korrelációs térképek (3.1–3.3. térkép) azt mutatják meg, mely kutatóponton van nagyobb különbség (illetve hasonlóság) a nyelvi hasonlósági viszonyokban, ha az eredeti (Fon.1) lejegyzéshez képest annak egyszerűsített formáiból készítünk dialektometriai elemzést.



3.1. térkép: A MNyA. finom fonetikai lejegyzéséből készített mátrix (Fon.1) és a mellékjeleket nem tartalmazó lejegyzésből készített mátrix (Fon.2) korrelációs térképe



3.2. térkép: A MNyA. finom fonetikai lejegyzéséből készített mátrix (Fon.1) és a fonológiai különbségek megőrzésével készített mátrix (Fon.3) korrelációs térképe



3.3. térkép: A MNyA. finom fonetikai lejegyzéséből készített mátrix (Fon.1) és a lexikai különbségekre érzékeny mátrix (Fon.4) korrelációs térképe

A 3.1. térképen a Fon.1 és a Fon.2 mátrix összevetésekor kapott korrelációs értékek szerepelnek. Minél jobban hasonlít a két elemzés eredménye egy kutatópont viszonylatában, a vöröshöz annál közelebbi, meleg (vörös, piros, téglaszín, narancs, sárga) színárnyalatot vesz föl, a színskálának megfelelően. A legalacsonyabb a korreláció Kéty esetében (r = 0,676), de még ez is elég erős pozitív korreláció. Ez azt jelenti, hogy a Fon.1 és a Fon.2 mátrix egyetlen kutatópont esetében sem tér el nagyon egymástól, sőt, a kutatópontok többségét tekintve az r értéke, kifejezetten magas: 0,9 fölötti. A mellékjelek elhagyása az eddigiek alapján legföljebb kis mértékben változtatja meg az egyes kutatópontok nyelvi hasonlóságának földrajzi mintázatát.

A fonetikai részleteket is figyelembe vevő (Fon.1) mátrix és a fonológiai különbségeket megőrző (Fon.3) között már általában véve is jelentősebbek a különbségek (3.2. térkép), szinte alig látunk pirosas árnyalatokat a térképen, néhány kutatópont esetében pedig csak igen gyenge pozitív korrelációt találunk. Csernátfalu (r = 0,235), Zágon (r = 0,355) és Halmágy (r = 0,445) esetében is igen alacsonyak az értékek. Ezen kutatópontok esetében tehát biztos, hogy jelentősebb eltérés mutatkozik a két dialektometriai elemzés eredménye között: a fonológiai jellegű lejegyzés nem képes a fonetikai részletek szintjén pontos lejegyzés alapján kirajzolódó nyelvi hasonlósági mintázatok megragadására.

A lexikai jellegű (Fon.4) mátrixszal való összevetésben (3.3. térkép) már csak a Nyugat-Dunántúlon látszanak piros, vagyis nagyobb korrelációra utaló árnyalatok, számtalan olyan kutatópont van a térképen, amelyek esetében a két mátrix csak közepesen erős korrelációt mutat (l. a zöldes és türkiz árnyalatokat). A legkisebb értékeket mutató (r < 0,6) kutatópontok listája a 3.1. táblázatban látható.

Kutatópont	Korreláció (r)
Csernátfalu	0,299
Zágon	0,314
Halmágy	0,408
Kakasd	0,477
Csíkrákos	0,483
Kéty	0,484
Oltszakadát	0,554
Vága	0,555
Hidas	0,561
Sókszelőce	0,576

3.1. táblázat: Együttjárás a Fon.1 és a Fon.4 mátrix között a legkevésbé korreláló kutatópontok esetében

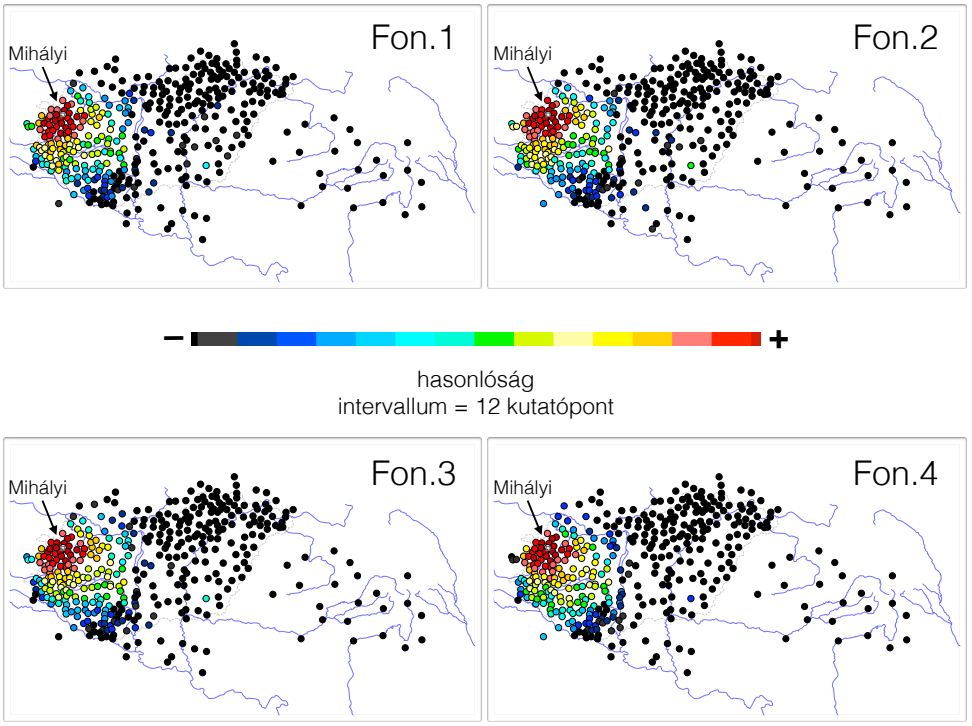
A legnagyobb eltérést a Fon.1 és a Fon.4 mátrix között erdélyi kutatópontok esetében tapasztalhatjuk, vagyis ezen kutatópontok esetében biztosan számottevően mást mutat az elemzés, ha fonetikai, illetve ha lexikai hangsúllyal vizsgálódunk. Annak magyarázatára, hogy éppen ezen kutatópontok esetében találunk nagyobb eltérést, feltehetően két szempontot is érdemes figyelembe vennünk: 1. elképzelhető, hogy ezeknek a településeknek távolabbi nyelvi kapcsolatai lehetnek, amelyeknek a hatása lexikai szinten már alig érvényesül; 2. a kutatóponthálózat Erdélyben igen ritka, így a közelebbi településekkel való nyelvi hasonlóság súlya csökken, a távolabbi kutatópontokkal való hasonlóság viszont fölerősödhet a fonetikailag érzékeny elemzésben. A Dél-Dunántúlra áttelepített bukovinaiak kutatópontjai szintén nagyobb eltérést mutatnak, de valójában ezek a kutatópontok csak látszólag dunántúliak, hiszen az érintett beszélőközösségek földrajzi helyzetében mindössze néhány évvel az adatfölvétel előtt következett be változás.

Vága és Sókszelőce esetében azonban nem szolgálhatunk az erdélyi kutatópontokéhoz hasonló magyarázattal. Mindkét kutatópont a palóc jellegzetességeket (leginkább az illabiális *ȧ* és a labiális hosszú *ā* meglétére gondolunk itt, vö. Sándor 2007) mutató terület nyugati szélén helyezkedik el. A fonetikai, illetve lexikai hangsúllyal készült elemzések gyenge korrelációja arra enged következtetni, hogy a két nyelvi szinten eltérőek ezen kutatópontok hasonlósági viszonyai: nem ugyanazokra a településekre hasonlítanak fonetikai, illetve lexikai szempontból.

A nagyobb eltérést mutató kutatópontok mellett érdemes azokra is kitérnünk, ahol nagyobb egyezést találunk. A nyugat-dunántúli települések mindhárom korrelációs térképen nagyobb együttjárást mutatnak, vagyis esetükben a fonetikailag érzékeny elemzés még a lexikai szintűtől sem tér el számottevően. Föltehető, hogy a dialektometriai elemzés eredményének hátterében településtörténeti jellegzetességek állnak. A Nyugat-Dunántúlra demográfiai, népmozgalmi szempontból a viszonylagos stabilitás jellemző az elmúlt sok száz évben, így a lakossággal együtt a nyelvváltozatok is viszonylagos állandóságot mutatnak.

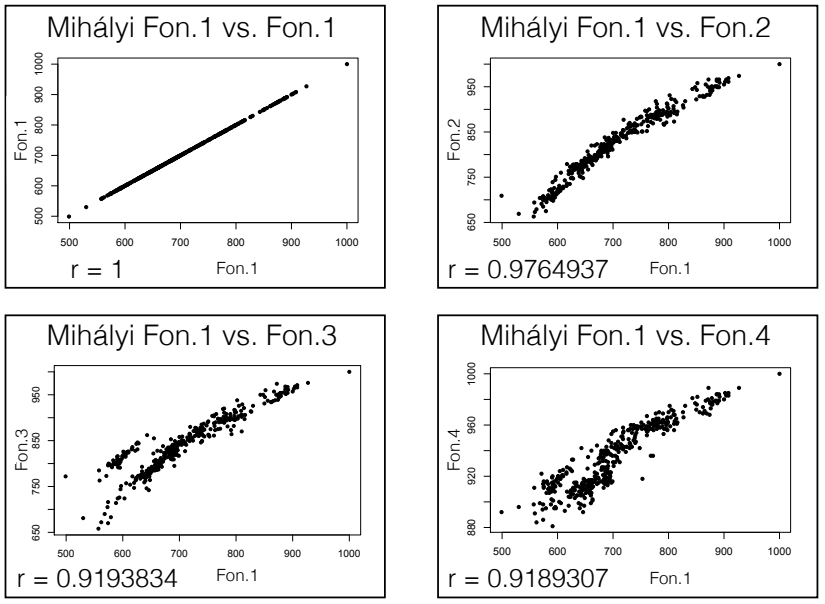
3.4. Egyes kutatópontok dialektometriai térképeinek elemzése

A korrelációs térképek alapján egyes kutatópontokra vonatkozóan kialakított hipotéziseinket egy-egy kutatópont különböző mátrixok alapján készített dialektometriai térképeinek összevetésével ellenőrizhetjük. A következőkben erre hozunk példákat.



3.2. ábra: Mihályi dialektometriai térképei különböző mátrixok alapján

A 3.2. ábrán Mihályi dialektometriai térképei láthatók. A térképen a vöröshöz közeli, meleg színárnyalatok jelzik a kiválasztott kutatóponttal nagyobb nyelvi hasonlóságot mutató településeket, a kisebb mértékű nyelvi hasonlóságot pedig hideg színek (zöld, kék), illetve a szürke és a fekete jelölik, a színskála szerint. A színek 12 kutatópontonként változnak, így, mivel a színskála véges, a kutatópontok többsége fekete. Mihályi a Nyugat-Dunántúlon található település, a korrelációs elemzés alapján azt várjuk, hogy ne legyen számottevő különbség a különböző mátrixok alapján készített hasonlósági térképei között. E kutatópont térsége a népmozgásokat tekintve az utóbbi évezredben stabilnak mondható, feltehetően ezzel is összefüggésben a vele szomszédos települések dialektusa igen hasonló. Viszonylag távol fekszik nyelvjárási törésvonalaktól. A négy térkép valóban szinte azonos, ezt a 3.3. ábrán látható pontdiagramok és a korrelációs értékek is alátámasztják. A korrelációs értékek és a pontdiagramok azt mutatják meg, milyen mértékben mutatnak egyezést Mihályi nyelvi hasonlósági viszonyai az egyszerűsített lejegyzés alapján készített mátrixok szerint az eredeti, finoman mellékjelezett lejegyzés alapján készített Fon.1 mátrixhoz képest.

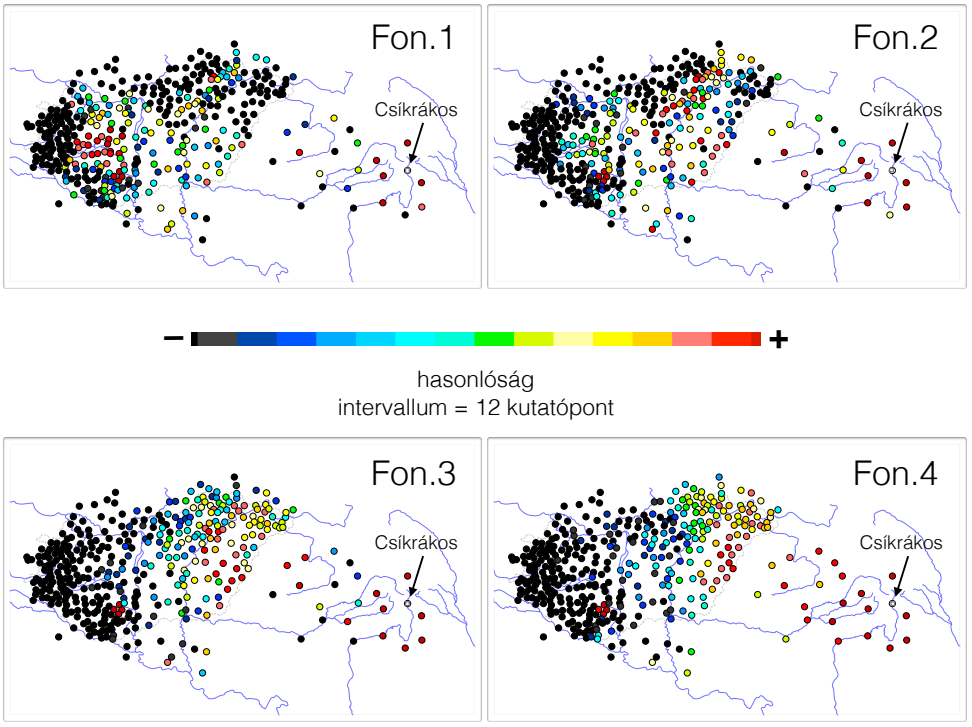


3.3. ábra: Korrelációk a Fon.1 és a többi mátrix között Mihályi esetében

A mátrixok közti korreláció, ha kizárólag Mihályi nyelvi hasonlósági viszonyait tekintjük, sokkal erősebb, mint az összes kutatópontot figyelembe vevő elemzés szerint. Az eredeti, finoman mellékjelezett lejegyzésből készített mátrix értékei 0.9 fölötti, tehát a tökéletes azonossághoz igen közeli mértékben korrelálnak a „fonológiai" (Fon.3) és a „lexikai" (Fon.4) jellegű mátrixban kapott számokkal. (Az itt látható értékek alapján alakul Mihályi ábrázolása a 3.3. fejezetben bemutatott korrelációs térképeken.)

Mihályiétól igen eltérő képet kaphatunk azonban, ha eredeti nyelvi környezetüktől évszázadokkal ezelőtt elszakadt települések, például nyelv(járás)szigetek (l. Szabó 1990, Péntek 2005) vagy a Dunántúl különböző részeiről a mai Székelyföldre áttelepült közösségek (l. Benkő 1967) eltérő mátrixok használata esetén kimutatható nyelvi hasonlósági viszonyai közt végzünk összevetéseket. Ilyen, a korrelációs térképek alapján is nagyobb eltérést mutató kutatópontok például Zágon és Csíkrákos. Érdemes azonban megnéznünk a szélsőségesen csekély korrelációs értékeket mutatók mellett más, szintén sziget helyzetű kutatópontokat is, ahol közepes mértékű a mátrixok közti együttjárás.

Csíkrákos nyelvi hasonlósági viszonyait különböző mátrixokon alapuló térképeken szemlélve (3.4. ábra) azt láthatjuk, hogy már a mellékjelek törlése szemmel látható hatással van a nyelvi hasonlósági viszonyok alakulására. A lejegyzés további, fonológiai szintűnek megfelelő egyszerűsítésével pedig a korábban nagyobb hasonlóságot mutató, pirosas árnyalatokkal előtűnő Balaton környéki kutatópontok egészen elsötétülnek, a térben közelebbi kutatópontok mutatnak egyre nagyobb hasonlóságot a kijelölt településsel. (A továbbra is piros dél-dunántúli pontok valójában „térben közeli" településeket jelölnek, hiszen lakosságuk a gyűjtés előtt alig tíz évvel települt át, így nyelvhasználatukra az új környezet kevésbé lehetett hatással.)



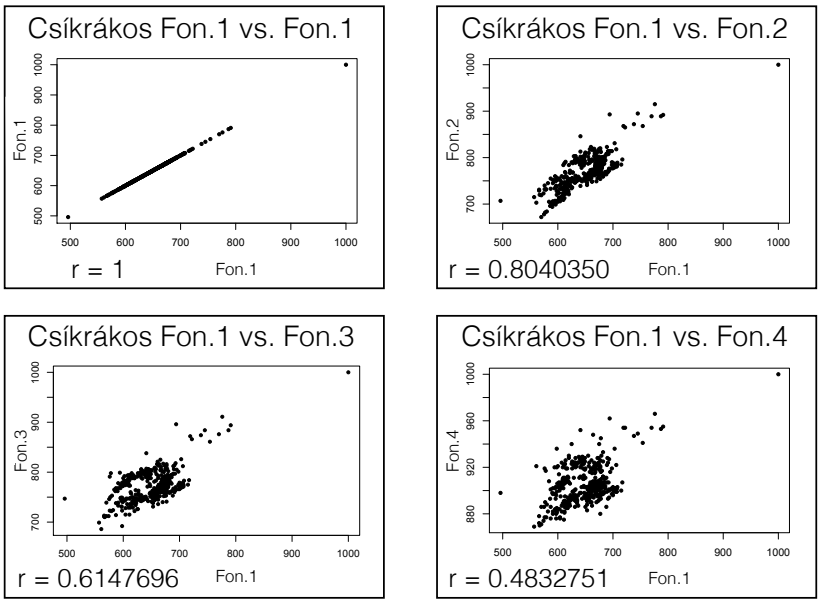
3.4. ábra: Csíkrákos dialektometriai térképei különböző mátrixok alapján

Csíkrákos esetében a korrelációs elemzés szerint (3.5. ábra) már a mellékjelek nélküli adatokból készített mátrix (Fon.2) is jelentősen eltér az eredetitől (Fon.1), a „fonológiai" pontosságú (Fon.3) és a „lexikai" (Fon.4) jellegű elemzés még ennél is kisebb, közepes korrelációt mutat az elsővel. Ha összevetjük az első pontdiagramokat, amelyeken a Fon.1 mátrixot önmagával korreláltattuk (3.3. és 3.5. ábra), látható, hogy Csíkrákos esetében a leghasonlóbb kutatópontok is sokkal kisebb mértékben mutatnak egyezést, mint Mihályi esetében.

Hasonló megállapításokat tehetünk egy másik keleti székely település, Zágon dialektometriai térképeiről (3.6. ábra), annyi különbséggel, hogy a mellékjelektől való eltekintés mintha inkább fölerősítené a nyugat-dunántúli, zalai nyelvjárásokkal való hasonlóság mértékét, miközben még jelentősebb a kontraszt a Fon.1 mátrix és a Fon.3, illetve a Fon.4 mátrix alapján készült térképek között. Ezzel összhangban az utolsó két mátrix gyengén korrelál az elsővel, ami arra enged következtetni, hogy egészen átrendeződnek a nyelvi hasonlósági viszonyok az elemzés hátterében álló jelenségek sajátságaitól függően (a mátrixok közti korreláció mértékéről Zágon és négy további kutatópont esetében lásd a 3.2. táblázatot).

A fonológiai különbségeket megőrző és a lexikai jellegű dialektometriai elemzés más települések esetében sem alkalmas térben és időben távoli nyelvi kapcsolatok megmutatására. Lássunk erre két példát olyan településekről, amelyeknek szoros kapcsolatát valamely távoli nyelvjárással korábbi kutatások már kimutatták. A klasszikus dialektológiai elemzések alapján Nagyhinddel (Benkő 1961) vagy más palóc településsel (Zelliger 1988) rokonítható kupuszinai nyelvjárás nyelvi hasonlósági viszonyai jelentősen megváltoznak, ha elemzésünkben eltekintünk a fonetikai részletektől (l. a 3.7. ábrát). A Szuhoggyal fonetikailag nagyobb mértékű egyezést mutató Kórógy (vö. N.

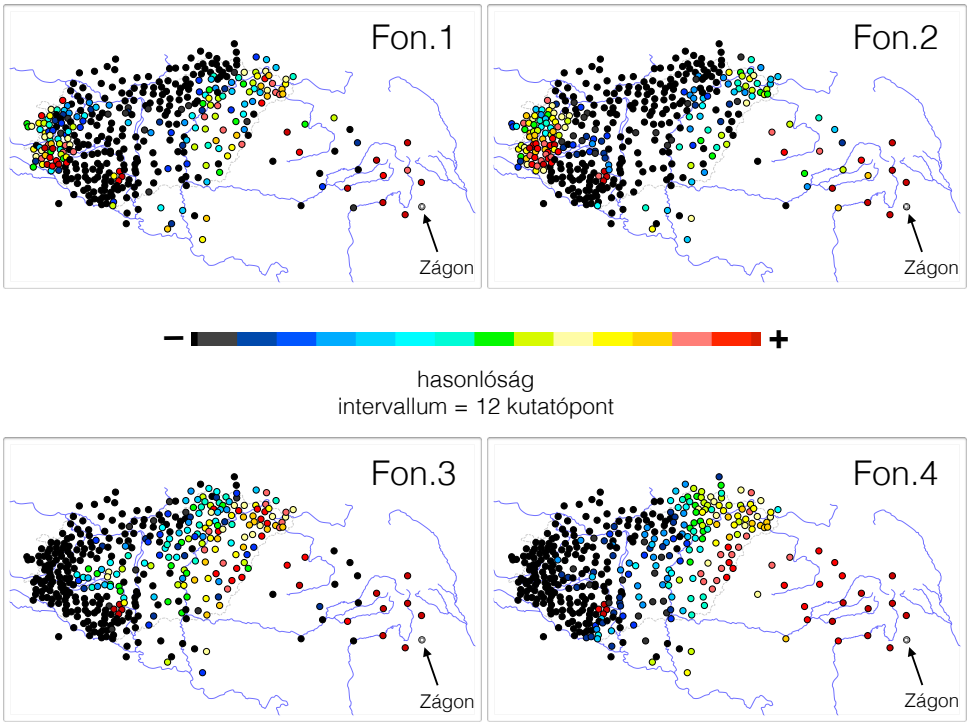
Fodor 2011) nyelvi hasonlósági viszonyai szintén megváltoznak, ha fonológiai vagy lexikai hangsúlyú összevetést végzünk (l. a 3.8. ábrát).



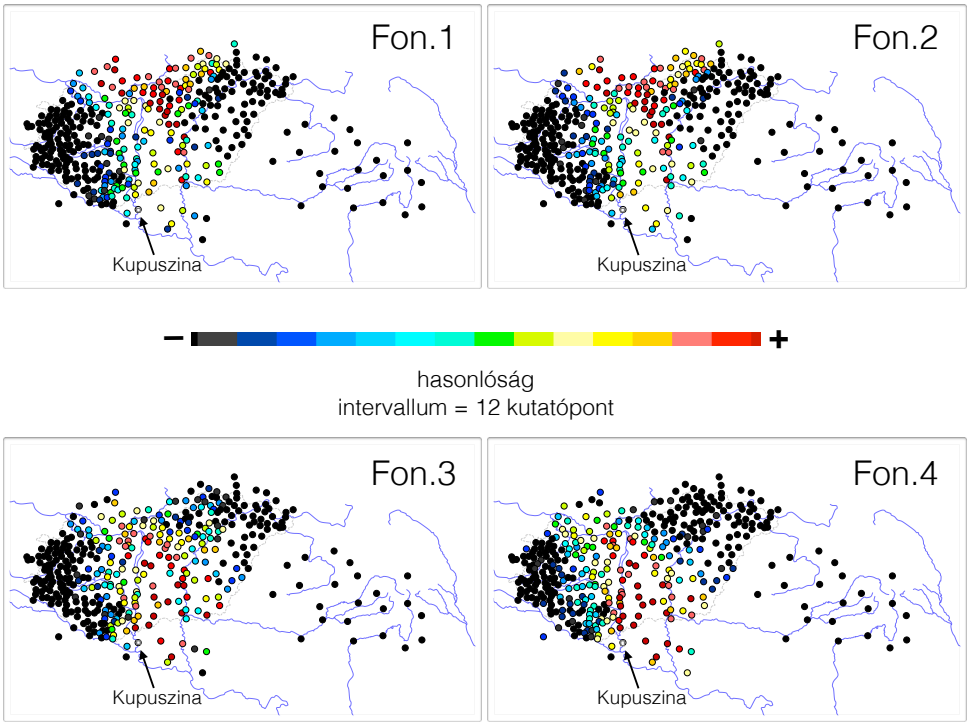
3.5. ábra: Mátrixok közötti korreláció Csíkrákos esetében

Kutatópont	Fon.1 vs. Fon.2	Fon.1 vs. Fon.3	Fon.1 vs. Fon.4
Zágon	0,756	0,355	0,314
Kupuszina	0,912	0,807	0,651
Kórógy	0,886	0,748	0,730
Szuhogy	0,906	0,864	0,747
Vága	0,891	0,747	0,555

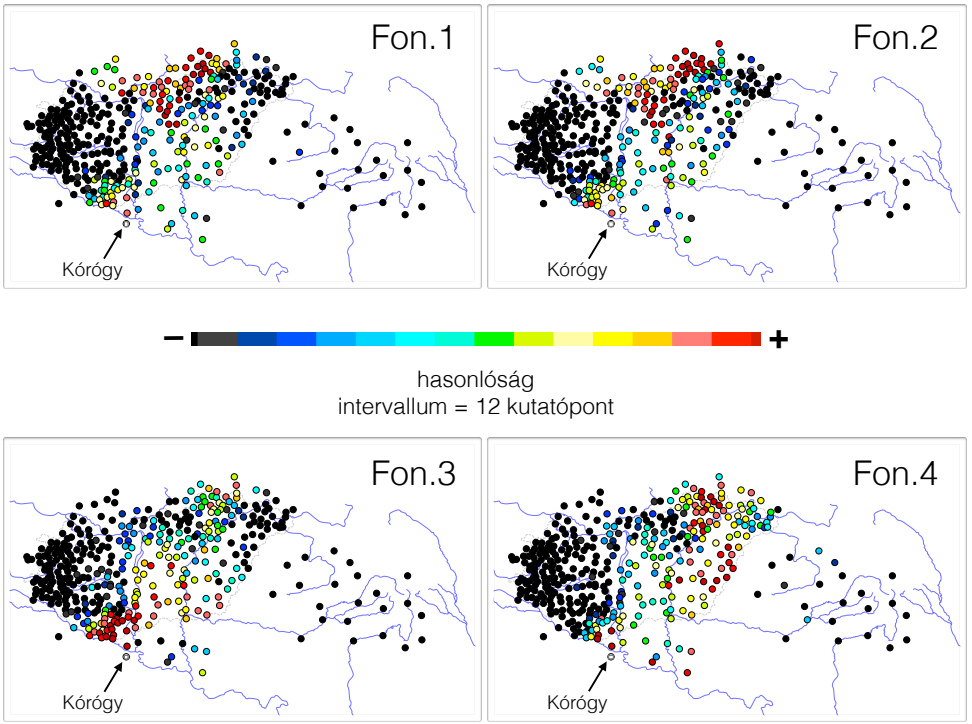
3.2. táblázat: Pearson-korreláció különböző mátrixok között Zágon, Kupuszina, Kórógy, Szuhogy és Vága esetében



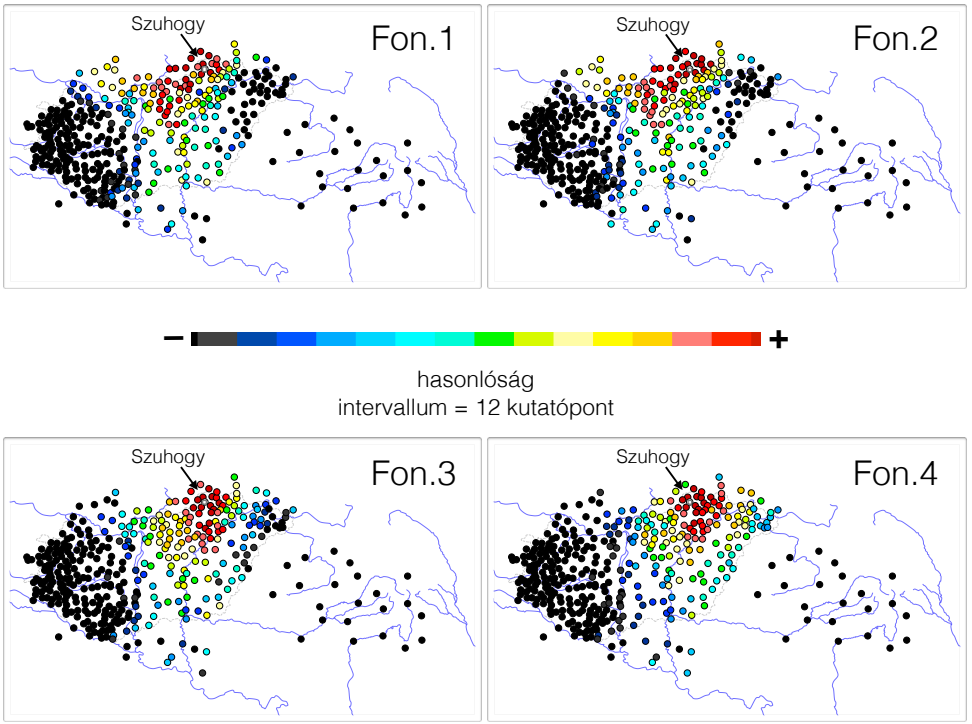
3.6. ábra: Zágon dialektometriai térképe különböző mátrixok alapján



3.7. ábra: Kupuszina dialektometriai térképei különböző mátrixok alapján



3.8. ábra: Kórógy dialektometriai térképei különböző mátrixok alapján



3.9. ábra: Szuhogy dialektometriai térképei különböző mátrixok alapján

A Kórógyhoz nyelvileg leginkább hasonló Szuhogy valamennyi térképén (3.9. ábra) olyan kutatópontokkal mutat leginkább nyelvi hasonlóságot, amelyek hozzá földrajzilag közel esnek. Vegyük észre azonban, hogy bár Szuhogy dialektometriai térképeit összevetve nincs olyan jelentős különbség a nagyobb hasonlóságot mutató kutatópontok térbeli elhelyezkedését illetően, mint Kórógy esetében, mégis jól látható, hogy a fonológiai és a lexikai jellegzetességeket fölerősítő elemzéseknél a nyelvi hasonlóság itt is a földrajzi közelség mentén alakul, a különböző mátrixok közti korreláció (3.2. táblázat) nem olyan erős, mint Mihályinál. A két település hasonlósági viszonyainak bizonyos fokú eltérésére az is magyarázatot adhat, hogy a Palócföld inkább egységes fonetikailag, mint lexikailag, így Szuhogy a fonetikai részletekre legkevésbé érzékeny lexikai mátrix alapján inkább mutathat egyezést a földrajzilag közeli, nem palóc nyelvjárású kutatópontokkal is. A palóc nyelvjárás határainak meghatározásában az illabiális *ȧ* és a labiális, hosszú *ā* megléte az alapvető szempont a dialektológiában (Sándor 2007), ez a hangtani sajátosság csak a Fon.1 és a Fon.2 mátrix hátterében lévő lejegyzésben van meg.

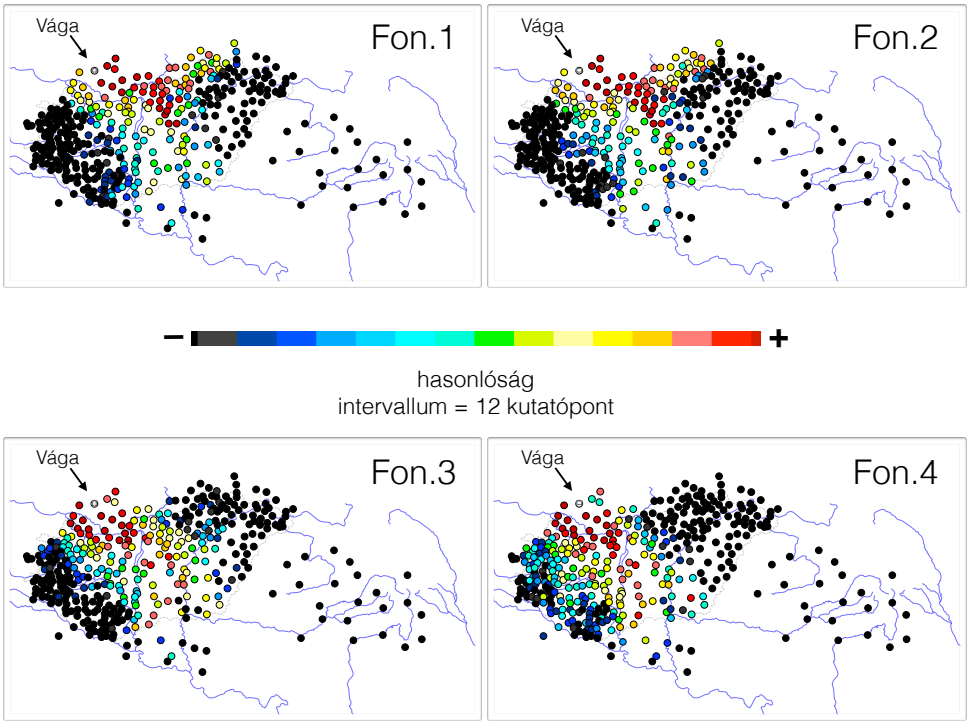
A korábbiakban bemutatott korrelációs térképek alapján a Palócföld nyugati szélén lévő kutatópontok nyelvi hasonlósági viszonyai eltérhetnek az elemzés hátterében lévő lejegyzés fonetikai finomságának függvényében. Ahogyan az a 3.2. táblázatban látható, Vága esetében például míg a Fon.1 és a Fon.2 mátrix között erős a korreláció, addig a Fon.1 és a Fon.3, illetve a Fon.1 és a Fon.4 mátrix között már csak közepes vagy annál is gyöngébb (3.10. ábra). A különbségek mutatnak némi hasonlóságot a Szuhogynál leírtakkal, annyi különbséggel, hogy Vága esetében a változás erőteljesebben mutatkozik meg, a kutatópont nyelvi hasonlósági súlypontja lexikai alapon vizsgálva ellentétes irányba mutat, mint a fonetikai mátrix alapján. Vága tehát fonetikai alapon nézve a palóc nyelvjárású kutatópontokhoz sorolható, lexikailag azonban Réte és Sókszelőce mellett leginkább a tőle délre fekvő csallóközi településekkel mutat nagyobb hasonlóságot. A Vágával leginkább hasonlóságot mutató Réte és Sókszelőce nyelvi viselkedése a korrelációs térképek alapján Vágáéhoz hasonlóan alakul.

A hangtani részletek figyelmen kívül hagyása, vagyis a lexikai jellegű különbségek súlyának növekedése az elemzésben a legtöbb esetben nem okoz alapvető változást a kutatópontok nyelvi hasonlósági viszonyaiban. Erre utal, hogy a legtöbb kutatópont esetében a Fon.1 mátrix még a Fon.4 mátrixszal is viszonylag erős pozitív korrelációt mutat (l. a 3.3. térképet), és erre utalnak a más nyelvjárási hatásoknak kevésbé kitett kutatópontok (pl. Mihályi, 3.2. ábra) hasonlósági mintázatai.

Az eredeti nyelvjárási környezetükből elvándorolt beszélőközösségek esetén azonban csak a fonetikailag érzékeny elemzés alkalmas az akár több évszázados nyelvjárási kapcsolatok megjelenítésére. Ezt mutatja a székelyföldi települések, Csíkrákos (3.4. ábra) és Zágon (3.6. ábra) példája, de a délvidéki nyelvjárásszigetek is, Kupuszina (3.7. ábra) és Kórógy (3.8. ábra).

Az új nyelvjárási környezetbe került települések esetében tehát igen eltérőek lehetnek a dialektometriai térképek, ha a fonetikai részletekre tekintettel lévő, illetve ha azokat figyelmen kívűl hagyó (fonológiai pontosságú) vagy szélsőségesen egyszerűsítő (lexikai jellegű) elemzést végzünk. Lehetséges magyarázat az eredmények alakulására, hogy a környező települések szókészletének elemei könnyen beépülhetnek, a hangtani hatásokkal szemben viszont időben ellenállóbbak a nyelvjárások. Ez azzal függhet össze, hogy a hangtani jellegzetességek rendszerszerűbbek, ezért nehezebben változnak, hiszen a hangtani rendszer elemei között erős kölcsönhatást feltételezhetünk, így egy elem változása az egész rendszer átrendeződését tenné szükségessé. Az itt elmondottakat megerősítik más nyelveken végzett dialektometriai elemzések is, ahol szintén

kimutattak különbségeket a lexikai és a fonetikai jelenségeken alapuló dialektometriai elemzések között (vö. Pickl et al. 2014), illetve az egyes nyelvi szintek (hangtani, morfológiai, lexikai szint) változékonysága tekintetében (Heeringa & Hinskens 2011). A hangtani és a lexikai alapú dialektometriai elemzés korrelációja francia és olasz nyelvszigetek esetében egybevág a magyar példákkal (Goebl 2008: 61-62, 2010: 444). Megjegyzendő azonban, hogy a magyar nyelvatlaszok, különösen a MNyA. és a RMNyA. inkább alkalmasak ilyen jellegű összevetésekre, mivel ezekben az adattárakban számos olyan település található, amelyek lakossága hosszabb-rövidebb ideje elvándorolt korábbi nyelvjárási környezetéből, sokat megőrizve azonban annak jellegzetességeiből.



3.10. ábra: Vága dialektometriai térképei különböző mátrixok alapján

Az eredmények arra engednek következtetni, hogy a környező nyelvjárások lexikai hatása nyelvjáráshatár közelében elhelyezkedő települések esetében is jobban érvényesülhet (l. pl. Vága). A lexikonban bekövetkező változást föltehetőleg fölerősítheti a közelben lévő más nyelvjárású, nagyobb települések nyelvi hatása is (vö. Pickl 2016).

A kvantitatív elemzés alapján valószínűsíthető, hogy a hangtani jelenségek közül a fonetikaiak (például a középső nyelvállású hosszú magánhangzók diftongálása vagy az egyes magánhangzók pontos minősége), és nem feltétlenül a fonológiai jellegűek bizonyulnak időtállóbbnak, a külső hatásokkal szemben ellenállóbbnak, tehát a rendszerszerűség szempontjából relevánsnak. Fölmerül, fontos-e, hogy a fonológia számot tudjon adni ezekről a jelenségekről, amelyek egy nyelvváltozat szempontjából lényegesnek minősülnek, vagy ezek a releváns kérdések elsősorban a fonetika, esetleg

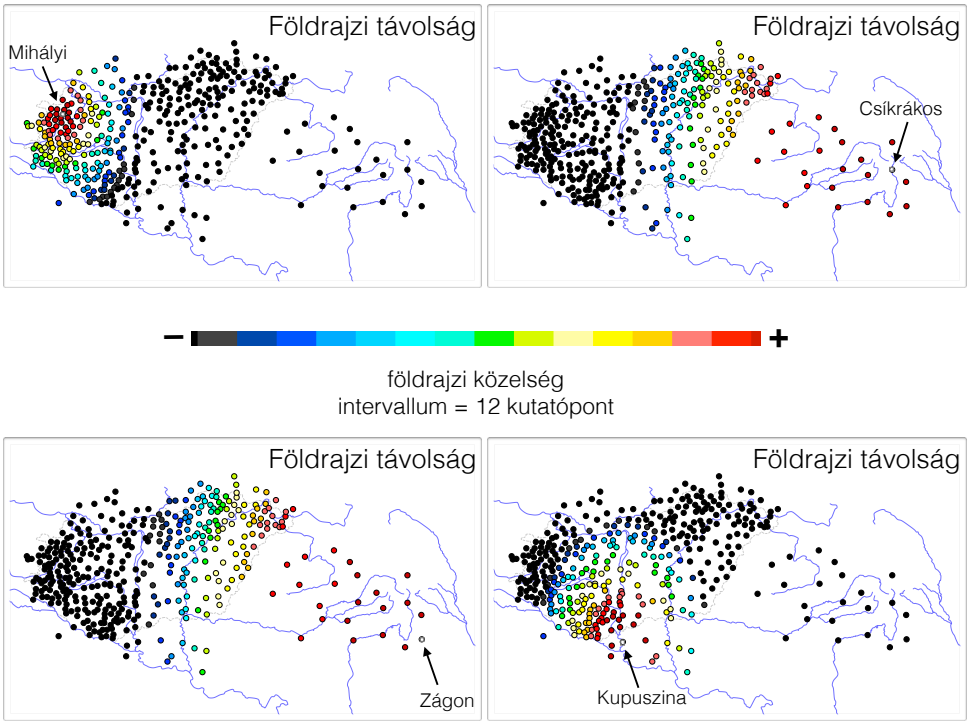
a fonetikai alapú fonológia (pl. az artikulációs fonológia, l. Browman & Goldstein 1992) területriumában maradnak.

3.5. Nyelvi hasonlóság és földrajzi távolság

Az egyes kutatópontok dialektometriai térképeinek összevetése alapján a fonetikai és a lexikai alapú mátrixok által megrajzolt nyelvi hasonlósági mintázatok, leginkább nyelvjárásszigetek esetében, számottevően eltérhetnek egymástól.

Amikor különböző mátrixokat vetünk össze, nem csak arra van lehetőségünk, hogy nyelvi mátrixokat korreláltassunk egymással. A kutatópontok közti földrajzi távolságról is készíthetünk mátrixot, amelyet összevethetünk a nyelvi távolsággal (Goebl 2005, Gooskins 2005). Sőt, az egyes kutatópontokról is készíthetünk olyan térképet, amely azt mutatja meg, mely kutatópontokhoz vannak földrajzilag a legközelebb, és melyektől vannak távolabb.

A 3.11. ábrán négy kutatópont, Mihályi, Csíkrákos, Zágon és Kupuszina földrajzi távolságmátrix alapján készített térképe látható. Minél közelebb van térben egy kutatópont a kijelölt településhez, annál melegebb (vörös, narancs, sárga) színt kap, minél távolabb van, annál hidegebb a színe (zöld, kék), a legtávolabbi pontok pedig szürkék, illetve feketék. A különböző (Fon.1, Fon.2, Fon.3 és Fon.4) nyelvi hasonlósági mátrixok alapján úgy látszik, hogy azon kutatópontok, amelyeknek hasonlósági mintázatai a Fon.1 mátrix alapján távolabbi helyekre mutatnak (pl. Zágon), a lexikai különbségekre érzékeny Fon.4 mátrixból készített dialektometriai térképen általában már csak a térben közeli kutatópontokkal mutatnak nagyobb hasonlóságot.



3.11. ábra: Mihályi, Csíkrákos, Zágon és Kupuszina földrajzi távolságmátrix alapján készített térképe

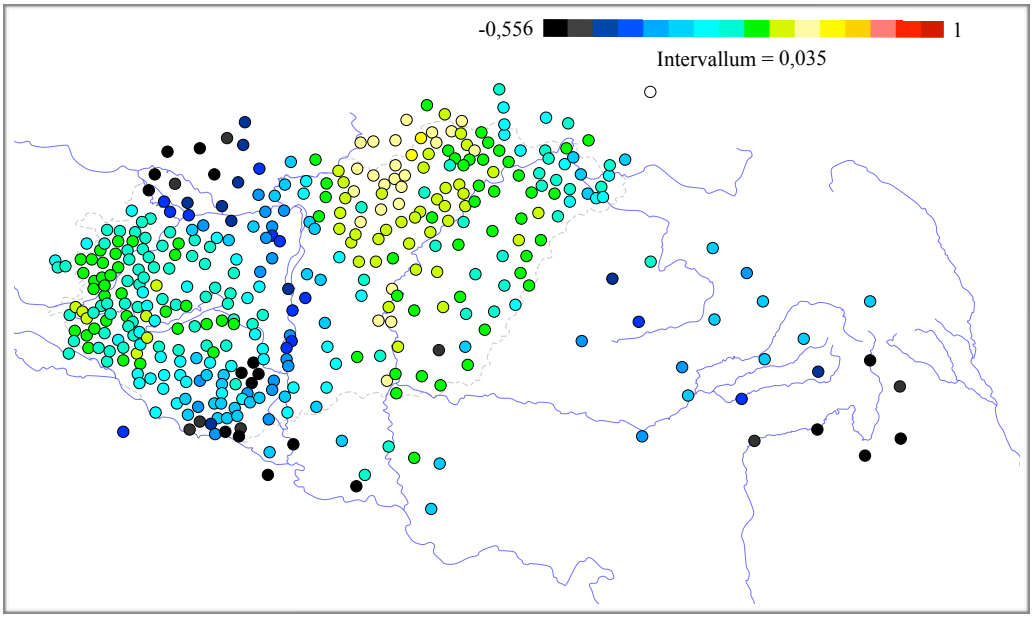
A 3.11. ábrán szereplő térképeket érdemes összevetnünk az egyes kutatópontok dialektometriai térképeivel. Mihályinál (3.2. ábra), ahol nincs érdemi különbség a négy dialektometriai mátrixszal készített térkép között, e négy korábbi térkép egyaránt hasonlít a kutatópontok földrajzi közelségét megmutató térképre. A másik három település esetében azonban (vö. 3.4., 3.6. és 3.7. ábra) a földrajzi közelséget ábrázoló térképek rendre a lexikai (Fon.4) mátrixból készített dialektometriai térképpel mutatnak leginkább hasonlóságot. A térképek összevetése és a korábbi elemzések eredményei alapján is arra következtethetünk, hogy a lexikai hasonlóság (Fon.4) inkább korrelál a földrajzi távolsággal, mint a fonetikai (Fon.1), illetve azon kutatópontok esetében, ahol nincs jelentős különbség a fonetikai és lexikai mátrix között, feltehetően már a fonetikai hasonlóság is nagyobb együttjárást mutat a földrajzi távolsággal.

A földrajzi közelség és a nyelvi hasonlóság összevetéséhez először is a mátrixok közötti korrelációt érdemes megvizsgálnunk. A négy nyelvi hasonlósági mátrix (Fon.1, Fon.2, Fon.3 és Fon.4) rendre a következők szerint korrelál a földrajzi közelséggel: Fon.1: r = 0,6422045, Fon.2: r = 0,615749, Fon.3: r = 0,8002051, Fon.4: r = 0,826772. A nyelvi hasonlóság természetesen a fonetikai részletek szintjén sem független a földrajzi távolságtól, hiszen a legtöbb kutatópont nem nyelv(járás)sziget, és nem nyelvjáráshatáron helyezkedik el, s így a hozzá földrajzilag közelebb lévő kutatópontokkal mutat nagyobb nyelvi hasonlóságot. A Fon.1 és a mellékjeleket nem, de az egyéb fonetikai jellegzetességeket tartalmazó Fon.2 mátrix is közepes pozitív korrelációt mutat a földrajzi távolsággal. A Fon.3 és a Fon.4 mátrixot ennél jóval nagyobb mértékű együttjárás jellemzi.

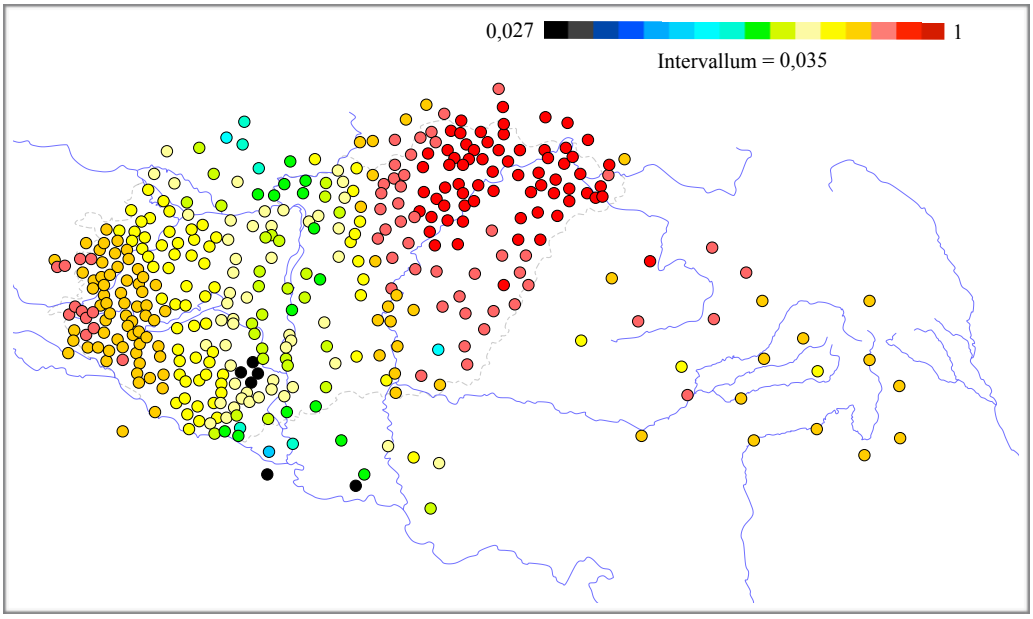
Ahhoz, hogy korrelációs térképeken is megvizsgálhassuk a fonetikai és a lexikai mátrix és a földrajzi közelség közti együttjárást, minden kutatópont esetében ki kell számítanunk a mátrixonkénti Pearson-korrelációt. Mivel a Fon.1 és a Fon.2 mátrix, illetve a Fon.3 és a Fon.4 mátrix igen hasonlóan viselkedik a Mantel-teszt eredménye alapján, elegendő a Fon.1 és a Fon.4, tehát a legapróbb fonetikai részleteket is figyelembe vevő és a lexikon szintjén releváns mátrixokon elvégeznünk az elemzést. A kutatópontonkénti értékeket a 3.4. és a 3.5.térkép szemlélteti.

Összevetve a két tréképet, a 3.4. térképen (Fon.1 vs. földrajzi távolság) egyáltalán nem találunk nagyobb korrelációra utaló pirosas árnyalatokat. A fonetikai hangsúllyal vizsgált nyelvi hasonlóság még a legmagasabb korrelációs értékeket mutató palócföldi kutatópontok esetében sem teljes mértékben a földrajzi távolság mentén alakul. Néhány kutatópont esetében pedig szinte nem is korrelál egymással a két mátrix (a korrelációs együttható értéke 0,07 és 0,28 közé esik). Ilyenek a bukovinai és moldvai áttelepültek falvai a Dél-Dunántúlon, de ilyen Kórógy, Zágon, Csíkrákos, Halmágy és Csernátfalu is, amelyeknél a nyelvi hasonlóság egészen távoli helyek felé mutat, illetve a ritka kutatópont-hálózat következtében sincsenek a közelükben más, hasonló nyelvjárású települések.

A 3.5. térképet (Fon.4 vs. földrajzi távolság) ezzel szemben meleg színek, vagyis magasabb korrelációs értékek jellemzik. A korábbi, egy-egy kutatópont dialektometriai térképeinek összevetésén alapuló megállapításunkat, miszerint a lexikai hasonlóság földrajzi mintázatai inkább mutatnak egyezést a földrajzi távolsággal, mint a fonetikaiak, a legtöbb kutatópont esetében szépen igazolja a térkép. Zágon és Csíkrákos esetében különösen nagy a két térkép közti különbség. Míg alig mutatnak valamelyest pozitív korrelációt a Fon.1-gyel való összevetésben, addig igen nagy (r = 0,88) a korrelációs együtthatójuk a 3.5. térképen (Halmágy és Csernátfalu esete is hasonló).



3.4. térkép: A MNyA. finom fonetikai lejegyzéséből készített mátrix (Fon.1) és a földrajzi közelség korrelációs térképe



3.5. térkép: A MNyA. adatainak lexikai különbségekre érzékeny változatából készített mátrix (Fon.4) és a földrajzi közelség korrelációs térképe

Kórógy esetében nincs jelentős változás, egyik térképen sem korrelál szinte egyáltalán a két mátrix. Az eredmény nem annyira meglepő, ha megnézzük a 3.8. ábrát, ahol látszik, hogy a lexikai hangsúlyú mátrix alapján is többségében térben távolabbi kutatópontokkal mutat nagyobb hasonlóságot.

A dél-dunántúli áttelepültek falvai esetében a 3.5. térképen közepes erősségű negatív korrelációt látunk, vagyis ezek a települések, érthető módon, inkább távolabbi kutatópontokkal mutatnak hasonlóságot, mint közeliekkel. Összhangban a többi székelyföldi kutatóponttal, ahol igen erős a pozitív korreláció a földrajzi távolsággal a lexikai jellegű mátrixszal való összevetés alapján.

A földrajzi távolsággal való összevetések alapján is igazolódni látszik az a korábbi megállapítás, hogy a gyorsabban változó lexikon inkább idomulhat a környező nyelvjárásokhoz, ezért inkább a földrajzi távolság mentén alakul. Ilyenek a MNyA. székelyföldi kutatópontjai. Nem igaz ez azonban minden sziget helyzetű településre, az elszigeteltebbek lexikai jellegzetességeiket is megőrizhetik. Ilyen település Kórógy, de feltehetően például a vajdasági Piros is, ahol egyik térképen sincs számottevően magas érték (Fon.1 r = 0,454, Fon.4 r = 0,384).

3.6. Összefoglalás

A fejezet a MNyA. sűrű kutatóponthálózatú informatizált térképlapjainak többszintű dialektometriai elemzésén keresztül járja körbe az egyes nyelvi szintekhez köthető, eltérő jellegű nyelvi hasonlósági mintázatokat, a nyelv(járás)szigetek speciális helyzetükből adódó jellegzetességeit, a földrajzi és a nyelvi távolság közti összefüggéseket. A fonetikailag érzékeny, a fonológiai pontosságú és a lexikai jellegű elemzés összevetéséhez összesen négy mátrixon végeztünk összevetéseket. Elsőként korrelációs elemzés és korrelációs térképek segítségével vizsgáltuk, mennyire változik meg a dialektometriai eredmény, ha az eredeti, finom fonetikai lejegyzést automatikusan egyszerűsítjük. Azt találtuk, hogy a mellékjelek hiánya még nem különösebben, a fonológiai szintű pontosságra, illetve a lexikai különbségekre való korlátozódás azonban már jelentősebben megváltoztatja az eredményeket, különösen egyes sziget helyzetű vagy nyelvjáráshatáron lévő települések esetében. Néhány konkrét település esetében összevetettük egymással a négy mátrix alapján kirajzolódó hasonlósági mintázatokat. A térképek alapján az imént említett speciális helyzetű kutatópontok esetében a különböző nyelvi szintekhez társítható hasonlóság földrajzi mintázatai akár jelentősen is eltérhetnek. A földrajzi és nyelvi távolság összevető elemzése alapján arra a következtetésre juthatunk, hogy míg a fonetikai hasonlóság kevésbé, a lexikai hasonlóság igen nagy mértékben a földrajzi távolság függvénye.

4. Integrált dialektometria – A magyar nyelvjárások atlasza és A romániai magyar nyelvjárások atlasza[5]

A magyar nyelvterület egésze nem vizsgálható csupán a MNyA. térképlapjai alapján, hiszen, ahogyan az előző fejezetben is láthattuk, a keleten meglehetősen gyér kutatóponthálózat hatással van a dialektometriai elemzésekre. Ahhoz tehát, hogy a teljes nyelvterületről átfogó elemzést készíthessünk, közös (integrált) kutatóponthálózaton kell együttesen vizsgálnunk a két nagyatlaszunk, a MNyA. és a RMNyA. adatait. A 4. fejezet fő kérdése, hogyan tudjuk ezt megvalósítani, mely térképlapok alkalmasak az elemzésre, milyen nehézségek merülhetnek föl az integrálás során. A két adattárról különböző módszerekkel készült integrált dialektometriai elemzéseket összevetve arra keressük a választ, melyik lehet az optimális módszer az adattárak egyesítésére. Külön alfejezetben lesz szó a köznyelv és a nyelvjárások közötti távolságot megmutató elemzésről.

4.1. Informatizált adattárak, az integrálás alapfeltételei

A nyelvatlaszok vagy egyéb nyelvjárási adattárak számítógépes feldolgozása során informatizált adatokat hozunk létre (az informatizálásról l. még a Bevezetést). Az informatizálás lényeges eleme, hogy minden, az adatra vonatkozó, rendelkezésünkre álló információt megőrzünk. A nyelvatlaszok esetében alapvető információn elsősorban az adat származási helyét, illetve jelentését (a térkép címszavát) érthetjük, de lehetnek még más, az adathoz kapcsolódó minősítések (pl. régies vagy ritka), megjegyzések is, amelyeknek szintén nem szabad elveszniük. Az eredeti adattárban szereplőkön túl, már a feldolgozás, adatelemzés során további attribútumok is kapcsolódhatnak az adatokhoz, például csoportosítási kódok, amelyeket szintén megőrizhetünk, és további kutatásokban is felhasználhatunk.

A Bihalboccsal informatizált adattárak azonban nem önmagukban állnak, hanem beépíthetők egy nagyobb rendszerbe, ahol – az azonos adatrögzítési elveknek is köszönhetően – összekapcsolódnak más adattárakkal. Az informatizálás révén így az egyes adatok értéke megnő, hiszen tágabb összefüggésekben szemlélhetjük azokat.

A különböző atlaszokból származó adatok integrálásához azonban feltétlenül szükségünk van integrált kutatópont-hálózatra is. Két (vagy több) adattár kutatóponthálózatának összekapcsolása természetesen nem egyszerűen csak annyit jelent, hogy az adattárakban kutatópontként szereplő valamennyi települést rárajzoljuk (automatizáltan rávetítjük) a térképre. Rendszerünknek pontosan tudnia kell, melyik pont melyik adattárnak eleme, illetve azt is, ha egyazon kutatópont egyszerre több adattárban is szerepel.

További fontos kritérium, hogy az alkalmazott nyelvészeti technológia az adatrögzítéshez használt magyar egyezményes hangjelölés rendszerét maximálisan támogassa, hiszen ez biztosítja, hogy a lejegyzést fonetikailag értelmezni tudjuk, és igény szerint egyszerűsíthessük, az eredeti, finoman mellékjelezett alapján új lejegyzési formákat hozva létre. A lejegyzés automatikus egyszerűsítése teszi lehetővé, hogy hatékonyabban tudjunk keresni az adatokban, illetve ezzel a módszerrel tudunk a dialektometriai elemzés során különböző nyelvi szinteknek többé-kevésbé megfelelő

[5] A fejezet az Atlaszintegrálás és kvantitatív adatelemzés (Vargha 2015), A magyar nyelvjárások atlasza és A romániai magyar nyelvjárások atlasza integrált dialektometriai elemzése (Kocsis–Vargha 2016) című tanulmányok és a Linguistic self-hatred and distance from standard Hungarian (Vargha 2016a) című előadás eredményeinek felhasználásával készült.

nyelvi hasonlósági mátrixokat létrehozni, mint ahogyan azt a 2. és 3. fejezetben is láthattuk (vö. továbbá Vargha 2015, Vargha, megjelenés előtt).

A MNyA. nyomtatásban megjelent térképlapjainak (összesen 1162 darab) informatizálása 2014-ben befejeződött. A RMNyA. 3297 térképlapjából összesen 1525 áll rendelkezésre ellenőrzött formában. Azonban nem minden informatizált térképlap integrálható. A legalapvetőbb kritérium, hogy legyen az egyik térképlapnak azonos címszavú párja a másik adatbázisban. Ez azonban még nem elégséges ahhoz, hogy az adott térképlapok a dialektometriai elemzésben is felhasználhatók legyenek, további feltétel, hogy a MNyA.-ból származó térképlapnak teljes kutatóponthálózatúnak kell lennie.

4.2. A különböző adattárakból származó adatok integrált elemzésének buktatói

A magyar nyelvjárási adattárak közül valamennyi a magyar egyezményes hangjelölési rendszer használatával készült, gondolhatjuk tehát, hogy a fent leírtak elegendők az adattárak integrált elemzésének biztosításához. Akkor azonban, ha az adatainkat nem önállóan szeretnénk vizsgálni, hanem kvantitatív módszerekkel aggregált adatokat szeretnénk létrehozni, újabb problémával szembesülhetünk: ugyan valamennyi adattár gyűjtői és lejegyzői a magyar egyezményes hangjelölést használták, ezt mégsem teljesen egyező gyakorlat mentén tették. Márpedig azokban a vizsgálatokban, ahol a fonetikus lejegyzés automatikus elemzésével dolgozunk, a hangjelölés gyakorlatában rejlő esetleges különbségek – megtévesztő módon – területi különbségekként jelenhetnek meg.

RMNyA.	darab	MNyA.	darab
o͡u̯	1706	o͡u̯	5230
o͡u̯	782	o͡u̯	2045
o͡u̯	778	o͡u̯	613
o͡u̯	696	o͡u̯	412
o͡u̯	605	o͡u̯	298
o͡u̯	566	o͡u̯	91
o͡u̯	477	o͡u̯	41
o͡u̯	396	o͡u̯	27
o͡u̯	240	o͡u̯	25
o͡u̯	210	o͡u̯	17

4.1. táblázat: az *óu*-féle diftongusok a MNyA. és a RMNyA. informatizált adataiban (forrás: Vargha 2015b)

A továbbiakban az *óu*-féle, középső nyelvállású, veláris labiális előtagú és felső nyelvállású veláris labiális utótagú diftongusok példáján mutatom be, milyen jellemző különbségek vannak a diftongusok lejegyzésében a két adattár között. Ilyen fajta diftongusok lejegyzésére a MNyA. 27, míg az RMNyA. 43 különböző jelölésmódot alkalmaz (természetesen igen eltérő gyakorisággal). A MNyA. esetében az öt leggyakoribb lejegyzési megoldás lefedi az összes eset (N = 8865) 97%-át, míg a RMNyA. esetében a tíz leggyakoribb jelölésmód is csak az 1516 felhasznált térképlap adataiban

előforduló esetek (N = 7186) 90%-át adja ki (lásd a 4.1. táblázatot). A hagyományosan emeléssel jelölt, nyomaték nélküli elemet rendszerünkben a magánhangzó fölé tett, fölfelé mutató nyíl fejezi ki (a lejegyzésről és annak számítógépes változatáról l. még az 1. fejezetet).

A kimutatás alapján egyértelmű, hogy a MNyA. a diftongusok elemei közti nyomatékeloszlás jelölésére kétféle grafikus módszert alkalmaz, az egyik az emelés (nyilacska), a másik az aláhurkolás. A RMNyA.-ban eléggé általánosan használt, a kevésbé nyomatékos elem alatt elhelyezett lebegő elem jel nem fordul elő az MNyA. adataiban. Az elemszámok arra engednek következtetni, hogy míg a RMNyA. igen változatosan oldja meg a diftongusok lejegyzését, mellékjelezését, addig a MNyA. egységesebb, kevesebb változatot használ, már a leggyakoribb öt jelölésmód szinte az összes előfordulást lefedi. Az alapvető különbségek más diftongusok jelölésmódjában is jelentkeznek, illetve a mellékjelek változatosabb használata a RMNyA.-ban általában is jellemző.

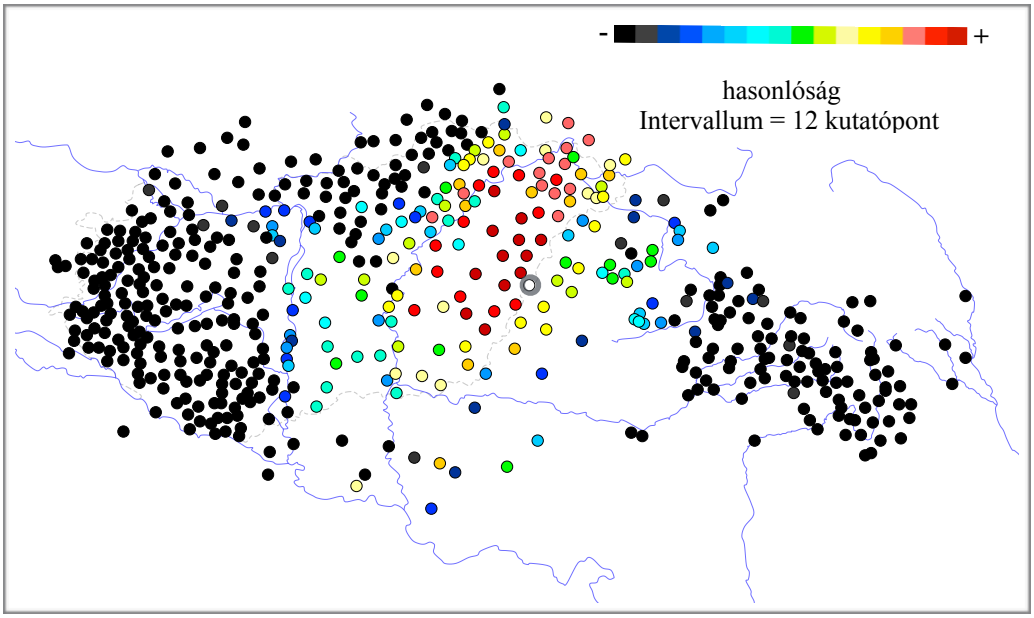
A diftongusok jelölésmódján túl vannak további, rendszerszerű különbségek is. A RMNyA.-ban például az úgynevezett széles ejtésű, a szokványosnál kissé nyíltabban ejtett hosszú magánhangzókat hosszúságjellel ellátott rövid megfelelőjük jelöli (pl. *ō, ȫ*). A MNyA. hangjelölési gyakorlatára ez nem jellemző, ott inkább a hosszú hangok nyíltabb mellékjeles változataival találkozhatunk, feltehetőleg hasonló funkcióval.

A két adattár egyesítésével végzett dialektometriai elemzést az eltérő lejegyzési gyakorlat több szempontból is torzíthatja. Mivel a MNyA.-nak megközelítőleg háromszor annyi kutatópontja van, mint a RMNyA.-nak, a MNyA. kutatópontjai az azonos stílusú lejegyzésnek köszönhetően könnyebben mutatnak nagyobb hasonlóságot a többi MNyA. kutatóponttal, mint a közeli, azonos nyelvjárástípusba tartozó RMNyA. kutatópontokkal. Ez a hatás, nevezzük adattárhatásnak, esetünkben két helyzetben is jelentkezhet: legjellemzőbb a két adattár határán, vagyis a magyar-román határ közelében, de érintheti a MNyA. valamennyi erdélyi kutatópontját is. A hatás fordítva (a RMNyA. felől nézve) kevésbé érzékelhető, mivel a RMNyA.-ban sokkal kevesebb a kutatópont, így azok a dialektometriai térképeken könnyebben mutatnak hasonlóságot a másik adattárból származó településekkel is.

Az adattárhatáshoz hasonló probléma más dialektometriai kutatások kapcsán is fölmerül, csak nem különböző adattárakhoz köthetően (lévén, hogy a magyar nyelvterületen kívül nincsenek még példák különböző adattárak integrálására), hanem a különböző terepmunkások eltérő lejegyzési gyakorlata és az ebből eredő területi különbségek kapcsán (Nerbonne–Kleiweg 2003, Mathussek 2016). Az itt elemzett adattárak közül a terepmunkások problémája a MNyA.-t érinthetné, a RMNyA.-t nem, hiszen azt egyetlen terepmunkás, Murádin László gyűjtötte. A MNyA. ugyan több terepmunkás részvételével, de jól összehangolt munkafolyamatban készült; a munkacsoport tagjai nagy hangsúlyt fektettek arra, hogy egységes legyen a lejegyzés. Olyannyira, hogy a zárt *ë* jelölésével kapcsolatos anomáliák (a hangjelölésben nem tükröződnek a keleti és nyugati nyelvjárások, illetve a Dunántúl és a Palócföld közti hangszín-realizációs különbségek) is következetesen jellemzik az egész gyűjtött anyagot (vö. Imre 1971: 272, Vargha 2013).

A két adattár tehát önmagában koherens, egymáshoz képest azonban nem feltétlenül azok. Ahhoz, hogy csökkentsük az eltérő lejegyzési szokásokból fakadó különbségeket, folyamodhatunk a lejegyzés egyszerűsítéséhez. Az előző fejezetekben erre már több példát is láthattunk. A MNyA. és a RMNyA. integrált adatbázisát elsőként itt is, mint a korábbi atlaszokat, a lejegyzés eredeti, finoman mellékjelezett formája alapján vizsgáljuk. Annak érdekében, hogy megvizsgáljuk, milyen hatása lehet a

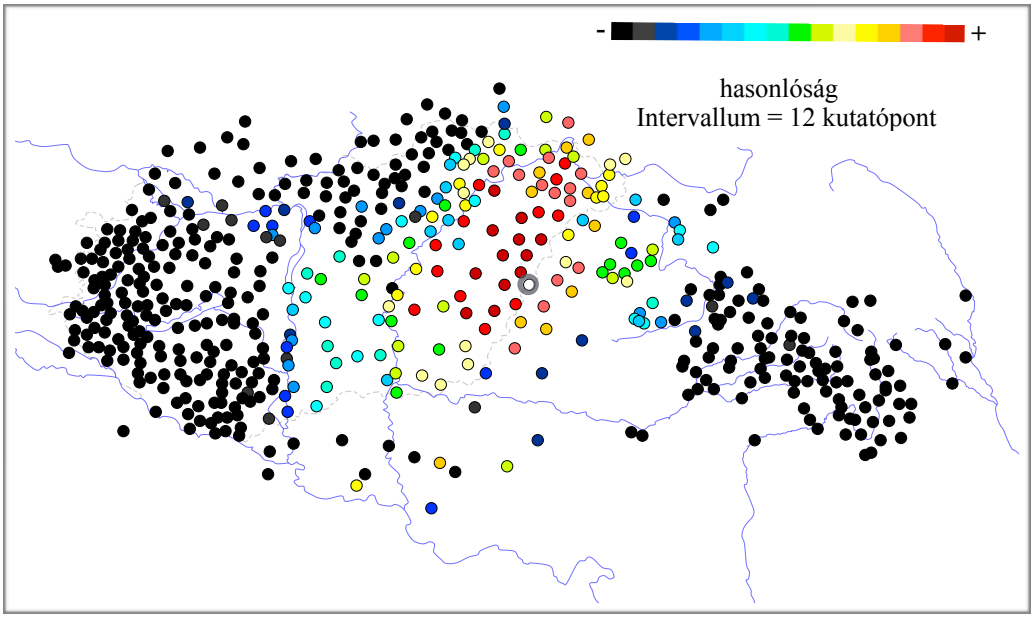
lejegyzés egyszerűsítésének az atlaszok közti adattárhatásra, létrehoztam a lejegyzésnek egy olyan változatát, amelyből a mellékjelek többsége hiányzik, de az időtartamot jelölők nem, illetve a hosszú hangokat azok hosszúságjellel ellátott rövid párjaival helyettesítettem. Ebben a fajta egyszerűsített lejegyzésben a magánhangzók időtartamának jelölése tulajdonképpen az IPA logikáját követi, abban a tekintetben, hogy az időtartamot sohasem alapjellel, hanem mindig mellékjellel fejezzük ki.



4.1. térkép: Ártánd dialektometriai térképe az eredeti, finoman mellékjelezett lejegyzés alapján

A 4.1. és a 4.2. térkép Ártánd nyelvi hasonlóságának földrajzi mintázatát mutatja a MNyA. és a RMNyA. 482–482 térképlapjának integrált dialektometriai elemzése alapján. A kijelölt kutatóponthoz (szürkével vastagon körvonalazva) viszonyítva, a színskálának megfelelően, meleg színek jelzik a nagyobb nyelvi hasonlóságot, a kevésbé hasonló kutatópontok hideg (zöld, kék), a legkevésbé hasonló kutatópontok fekete színben látszanak. Egy szín 12 kutatópontot fed le, így összesen 180 kutatópont színekkel differenciált megjelenítésére van lehetőség. A többi kutatópont egyaránt fekete színt kap. A 4.1. térkép az eredeti, mellékjeles lejegyzés, a 4.2. térkép a mellékjeleket csak részlegesen tartalmazó, módosított lejegyzés alapján készült.

Ártánd a MNyA. kutatópontja, az országhatár közelében fekszik. Nyelvjárására jellemző a záródó diftongusok használata, amelyek jelölésmódja (feltehetőleg nem kizárólagosan) felelős lehet azért, hogy a 4.1. ábrán az országhatár, amely esetünkben a két integrált adattárat is egymástól földrajzilag elhatárolja, nyelvjárási törésvonalnak tűnhet: keleti irányban nem, csak nyugatra és északra, a MNyA. kutatópontjai között látunk nagyobb hasonlóságot jelző pirosas árnyalatokat. A lejegyzés leegyszerűsített változata alapján, megszüntetve a diftongusok nyomatékeloszlásának jelölésében rejlő különbségeket is, az adattárak közti határ érzékelhetően valamivel kevésbé markánsan jelentkezik. Nem csak az azonos adattárban lévő, nyugati, hanem a közelben lévő keleti, már a RMNyA.-hoz tartozó kutatópontok közül négy is téglaszínben látszik, már nem határolódnak el annyira élesen a kiválasztott kutatóponttól.

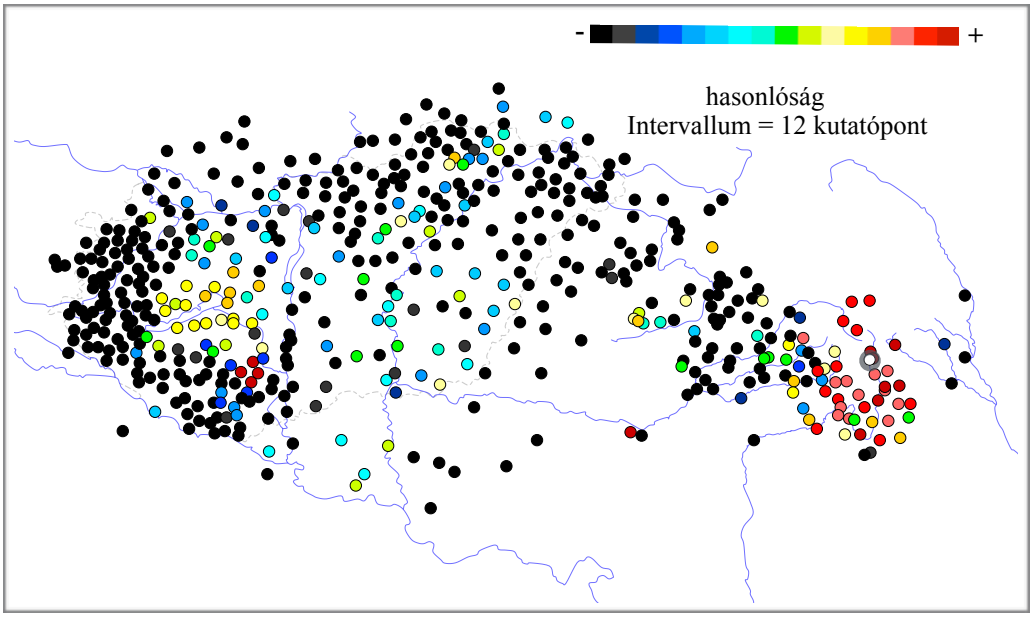


4.2. térkép: Ártánd dialektometriai térképe a módosított, mellékjeleket csak részlegesen tartalmazó lejegyzés alapján

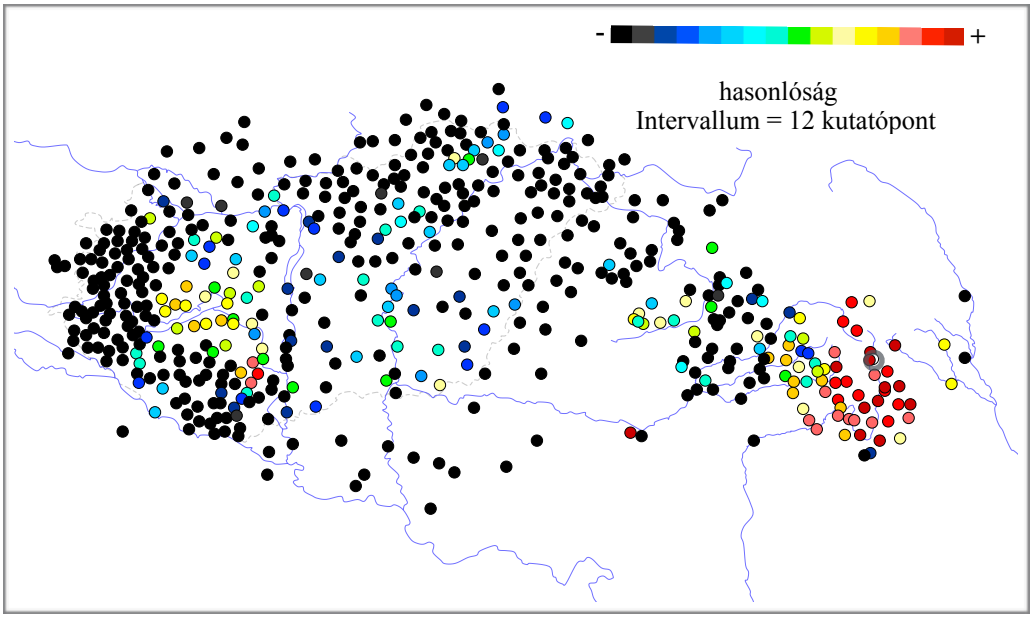
Érdemes részletesebben megnéznünk még olyan kutatópontpárokat, amelyek térben, és feltehetőleg nyelvjárásukban is, igen közel vannak egymáshoz, de különböző adattárba tartoznak. A 3. fejezetben, a MNyA. adatain végzett elemzésekben láttuk, hogy Csíkrákos a lejegyzés finoman mellékjelezett, eredeti formájának dialektometriai elemzése alapján nagyobb mértékű hasonlóságot mutat Balaton környéki kutatópontokkal (3.4. ábra). Mivel a MNyA. erdélyi részén nagyon gyér a kutatóponthálózat, föltehető, hogy a RMNyA.-val való integrálás hatására (több lesz a földrajzilag és nyelvjárásilag közeli kutatópont a térképen) kevésbé markánsan látszik majd a távolabbi kutatópontokkal való nyelvi kapcsolat. Csíkrákos térképét a vele szomszédos kutatópontéval, Vacsárcsiéval érdemes összevetnünk, amely már a RMNyA.-hoz tartozik (l. a 4.3. és a 4.4. térképet).

A két térkép rendkívül hasonló. Lényegében annyi különbséget figyelhetünk meg, hogy Csíkrákos esetében valamivel nagyobb hasonlóságot mutatnak a Balaton környéki kutatópontok, míg Vacsárcsinál a háromszéki, illetve az udvarhelyszéki kutatópontok hasonlítanak jobban. A két település hasonlósági viszonyai közt igen erős a korreláció (Pearson r = 0,903).

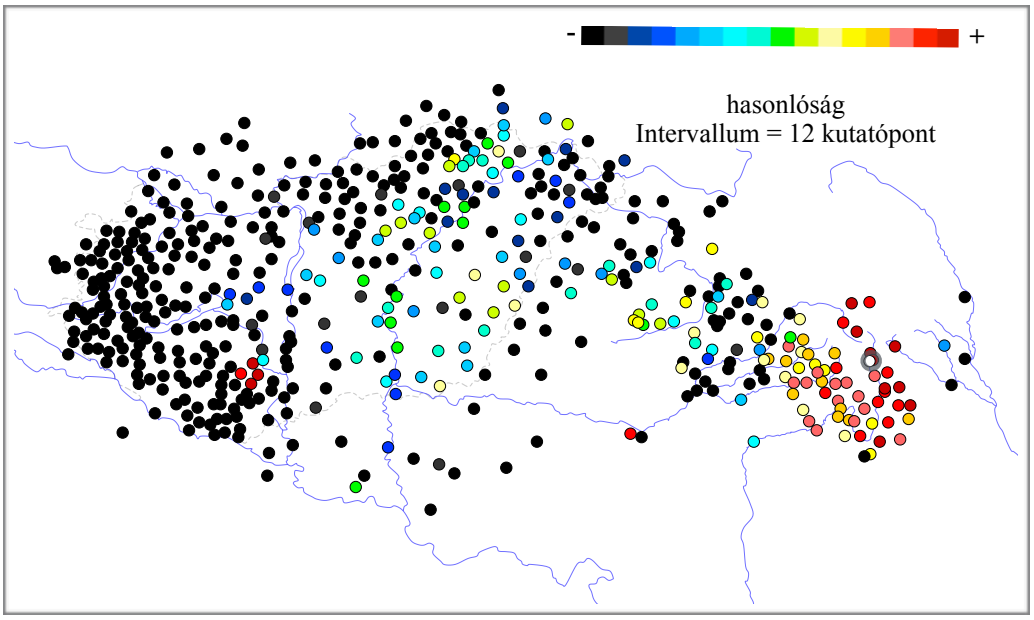
A lejegyzés egyszerűsítésével mindkét település esetében elhalványul a dunántúli kapcsolat, összhangban azzal, ahogy az előző fejezetben is láttuk Csíkrákos esetében (3.4. ábra). A két dialektometriai térkép szinte teljesen azonos, csak néhány kutatópontnál figyelhetünk meg eltérést a hasonlóság mértékét illetően. A korreláció itt erősebb a két kutatópont hasonlósági viszonyai között (Pearson r = 0,977), igen közel áll a teljes egyezéshez. A lejegyzés kisebb mértékű egyszerűsítésével tehát, ahogyan azt Csíkrákos vonatkozásban már korábban is megállapítottuk, a távolabbi településekkel való hasonlóság elhalványul, eltűnnek azonban a feltehetőleg az adattárhatás miatt jelentkező különbségek is, vagyis a két adattár szomszédos kutatópontjainak nyelvi hasonlósága erősebb korrelációt mutat.



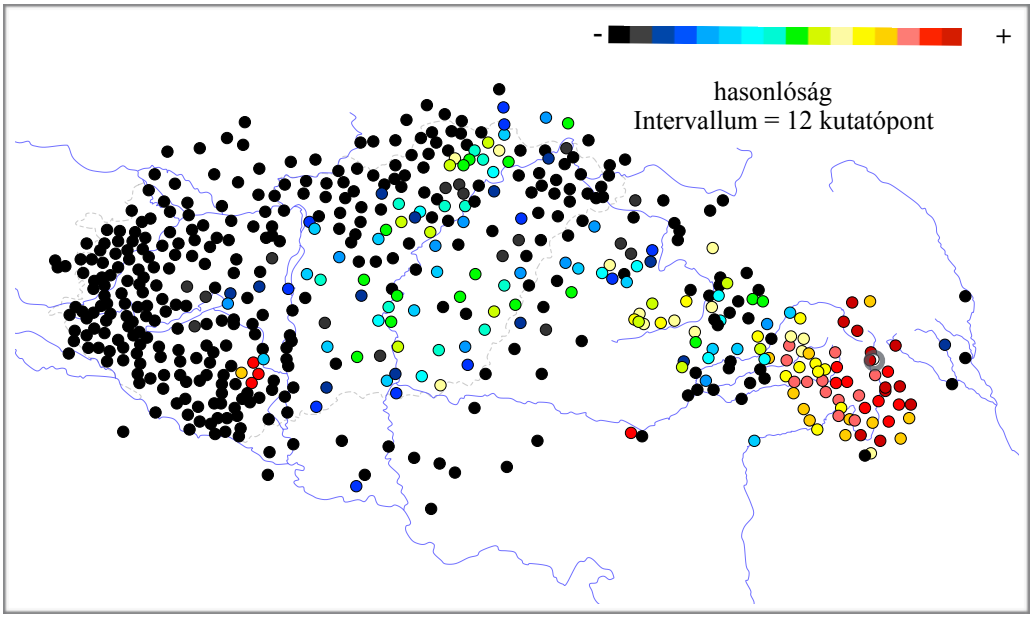
4.3. térkép: Csíkrákos dialektometriai térképe az eredeti, finoman mellékjelezett lejegyzés alapján



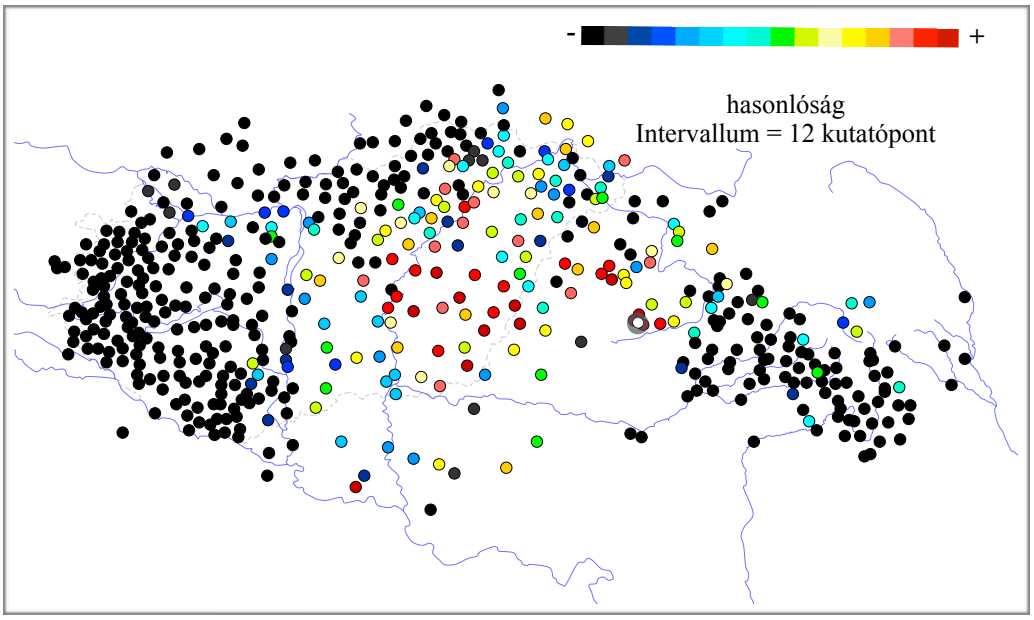
4.4. térkép: Vacsárcsi dialektometriai térképe az eredeti, finoman mellékjelezett lejegyzés alapján



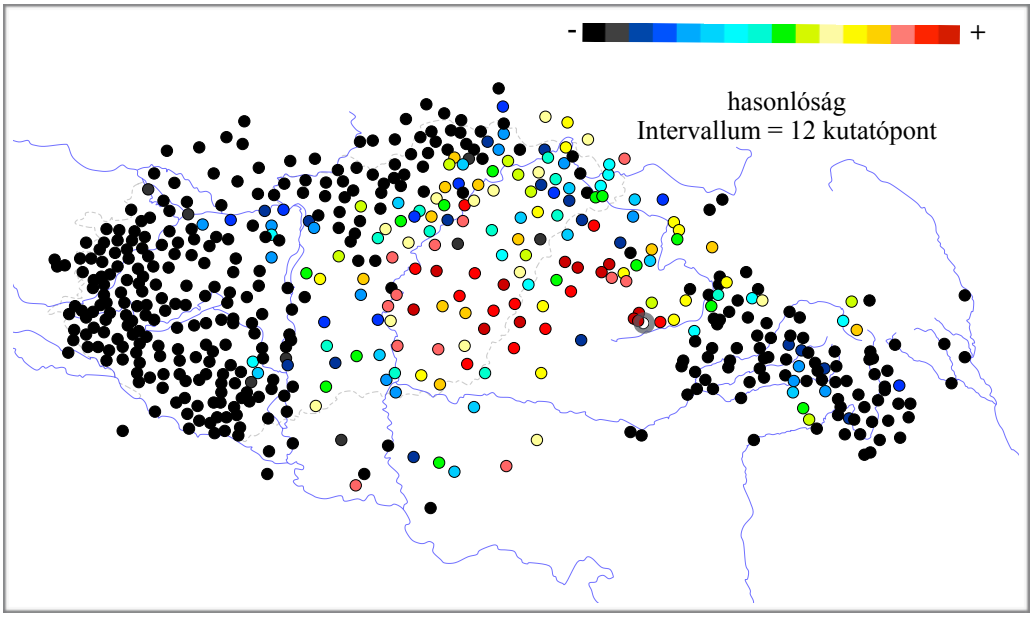
4.5. térkép: Csíkrákos dialektometriai térképe az egyszerűsített lejegyzés alapján



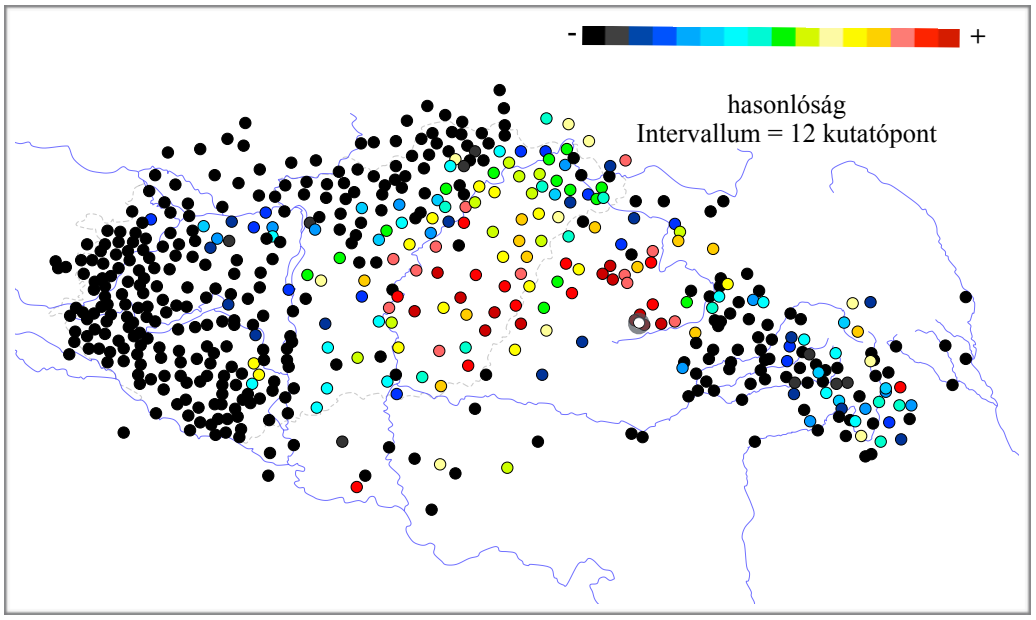
4.6. térkép: Vacsárcsi dialektometriai térképe az egyszerűsített lejegyzés alapján



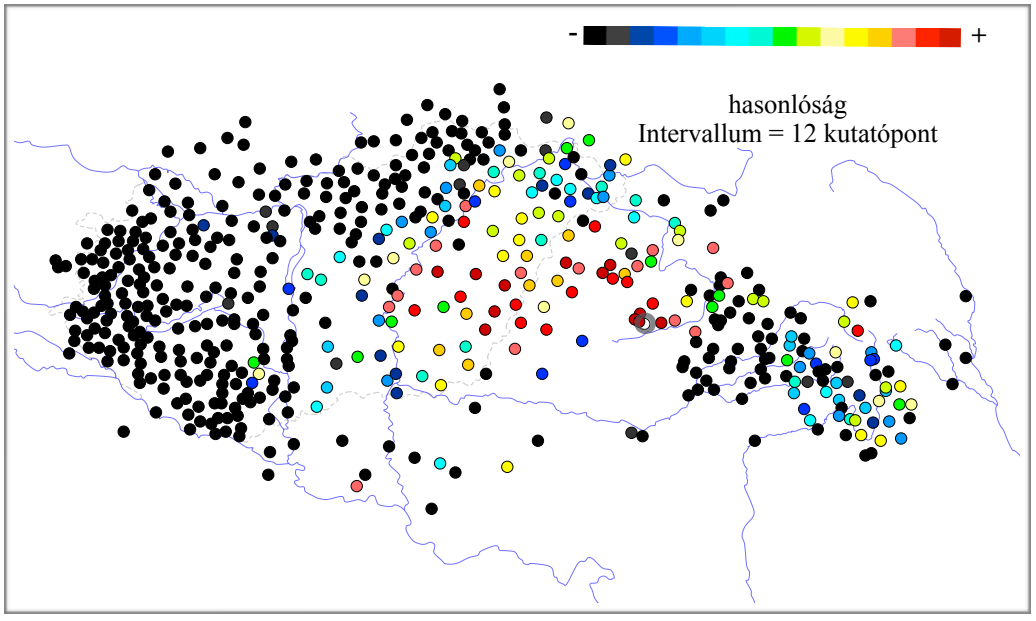
4.7. térkép: Magyarvalkó dialektometriai térképe az eredeti, finoman mellékjelezett lejegyzés alapján



4.8. térkép: Magyargyerőmonostor dialektometriai térképe az eredeti, finoman mellékjelezett lejegyzés alapján



4.9. térkép: Magyarvalkó dialektometriai térképe a módosított lejegyzés alapján



4.10. térkép: Magyargyerőmonostor dialektometriai térképe a módosított lejegyzés alapján

Két szomszédos kalotaszegi település, Magyarvalkó (MNyA.) és Magyargyerőmonostor (RMNyA.) között is végezhetünk hasonló összevetést. Az eredeti, mellékjeles lejegyzés alapján készült dialektometriai elemzés eredményét a 4.7. és a 4.8. térkép mutatja. A két település hasonlósági viszonyai szinte teljesen egyező mintázatot rajzolnak elénk. Mindkét kutatópont nyelvi hasonlósági súlypontja inkább nyugat felé mutat, vagyis inkább a MNyA. kutatópontjaira hasonlítanak. A korreláció a két kutatópont hasonlósági viszonyai között igen közeli a maximálishoz (Pearson r = 0,968). A lejegyzés egyszerűsítésének hatására nem sokat változik Magyarvalkó és Magyargyerőmonostor dialektometriai térképe, a legszembetűnőbb változás, hogy mindkét település esetében valamivel nagyobb lesz a hasonlóság két keletebbre lévő kutatóponttal, Középlakkal és Mérával. A hasonlósági viszonyok közti korreláció sem változik számottevően, bár a Csíkrákos és Vacsárcsi esetében megfigyeltekkel azonos irányban (Pearson r = 0,969).

A fentieket összefoglalva elmondhatjuk, hogy különböző adattárakból származó térképlapok integrálásakor számolnunk kell az adattárhatással, vagyis a lejegyzési gyakorlat különbségeiből adódó eltérésekkel. A lejegyzési stílus azonossága mesterségesen növelheti az azonos adattárból származó kutatópontok közti hasonlóságot. A hatás leginkább a több kutatóponttal rendelkező adattárat, esetünkben a MNyA.-t érinti. Mivel a RMNyA. kutatópontjaiból kevesebb van, azok könnyebben mutatnak hasonlóságot a MNyA. kutatópontjaival is.

Az adattárhatás leginkább a két atlasz határán jelentkezik, amely egybeesik az országhatárral. A határhoz közeli MNyA. kutatópontok kevésbé mutatnak hasonlóságot a határ túloldalán lévő RMNyA. kutatópontokkal (l. Ártánd 4.1. és 4.2. térképeit). A túloldalon lévő RMNyA. pontok kiválasztásakor azonban jellemzően a túloldalon is találunk bőven pirosas árnyalatú, nagyobb hasonlóságot mutató kutatópontokat. A határtól távoli, szomszédos MNyA. és RMNyA. kutatópontok hasonlósági viszonyaiban is lehetnek a lejegyzési szokásokból adódó különbségek (l. Csíkrákos és Vacsárcsi térképeit, 4.3.–4.6.).

Mivel az adattárhatás legfőbb oka az eltérő mellékjelezés, a lejegyzés egyszerűsítésével csökkenthető, ahogyan azt a fenti példákon láttuk. Lehetséges azonban, hogy az egyszerűsítés sem szüntet meg minden, lejegyzési szokásokkal összefüggő különbséget, míg esetleg nyelvileg pertinens részleteket eltüntethet (Csíkrákos és Vacsárcsi például csak az eredeti lejegyzés alapján mutat nagyobb mértékű hasonlóságot a Balaton-környéki kutatópontokkal). Az automatikus egyszerűsítés tehát előrelépést jelent az adattárhatás csökkentésében, föltehető azonban, hogy sokkal koherensebb integrált elemzést készíthetnénk olyan eljárás kidolgozásával, amely pontosan figyelembe veszi, és módszeresen megszünteti a hangjelölési gyakorlat eltéréseit. Ehhez azonban a teljes informatizált adattárban át kell tekinteni valamennyi alapjel-mellékjel kombinációt, és a különböző jelölésmódokat egyenként összehangolni. A lejegyzés ily módú összehangolása megvalósítható ugyan, de igen időigényes. Különösen addig, amíg a RMNyA. adattára nem teljes, így a benne lévő jelölésmódok nem tekinthetők át a maguk teljességében, hatékonysága miatt érdemesebb az automatikus egyszerűsítést választanunk.

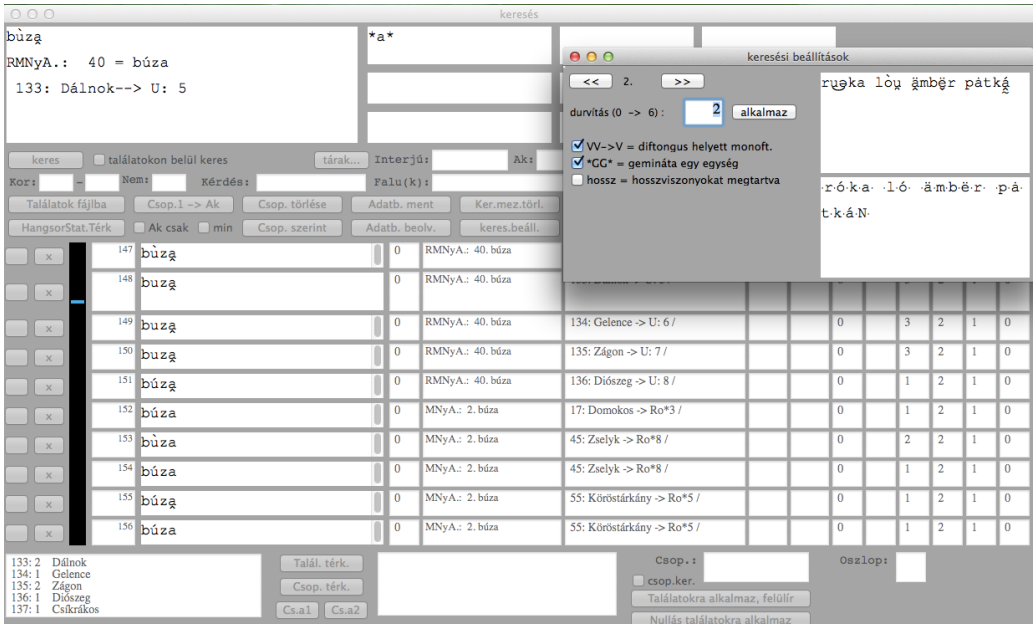
Az egyes kutatópontok nyelvi hasonlósági mintázatait vizsgáló dialektometriai térképek, valamint az egyes kutatópontok különböző mátrixok alapján kirajzolódó nyelvi hasonlósági viszonyait összevető korrelációs térképek esetében az adattárhatás még csak legföljebb egy-egy település esetében okozhat problémákat, akkor is látszólag csekélyeket. Akkor azonban, ha a mátrixokat további statisztikai elemzésnek szeretnénk alávetni, például a nyelvjárások automatikus felosztása céljából, már

okozhat problémákat, ha a különböző adattárak közt a lejegyzési jellegzetességek miatt keletkezik nyelvjáráshatár. A nyelvjárások automatikus felosztásával, a különböző mátrixok ilyen célú felhasználásával az 5. fejezet foglalkozik, a MNyA. és a RMNyA. integrált térképeit l. az 5.5. fejezetben.

4.3. A MNyA. és a RMNyA. integrált elemzése az adatok csoportosítása alapján

Az első fejezetben két dialektometriai elemzési módszerrel ismerkedhettünk meg. Az egyik, a jelen munkában általánosabban alkalmazott módszer az adatok automatikus összevetésén, a másik – a korábban kifejlesztett salzburgi módszer – az adatok előzetes, kutatók által végzett csoportosításán alapul. Mivel a magyar nyelvatlaszokból már több mint egymillió adatot számláló informatizált korpusszal rendelkezünk, ezért jóval hatékonyabb, így kézenfekvőbb a magyar nyelvjárások dialektometriai elemzésére az automatikus összevetést alkalmazni. Lehetnek azonban előnyei az adatok csoportosításának is, például két adattár integrálásakor kutatói csoportosítás esetén várhatóan nem, vagy csak igen elenyésző mértékben érvényesül az adattárhatás.

A Bihalbocsban lehetőségünk van egy-egy térképlap adatairól eltérő szempontok érvényesítésével összesen akár ötféle – Goebl munkatérképeinek megfelelő – csoportosítást készíteni. Az adatokban való keresést, így magát a csoportosítás folyamatát a lejegyzés egyszerűsítésének lehetősége, annak tetszőleges beállítása teszi könnyebbé, így hatékonyabbá (l. 4.1. ábra). A csoportosítás során mindig egyetlen szempontot alkalmazunk. A példában szereplő *búza* adatokból háromféle szempont mentén háromféle csoportosítást, vagyis három munkatérképet hoztunk létre.[6]



4.1. ábra: Csoportosított munkatérképek létrehozása a Bihalbocsban

[6] A munkatérképek közül néhány – köztük a *búza* három térképe is – elérhető a MNyA. és a RMNyA. integrált dialektometriai elemzésének eredményeit interaktív dialektometriai térképek formájában bemutató honlapon: http://bihalbocs.hu/mnyarmnya/intterk.html.

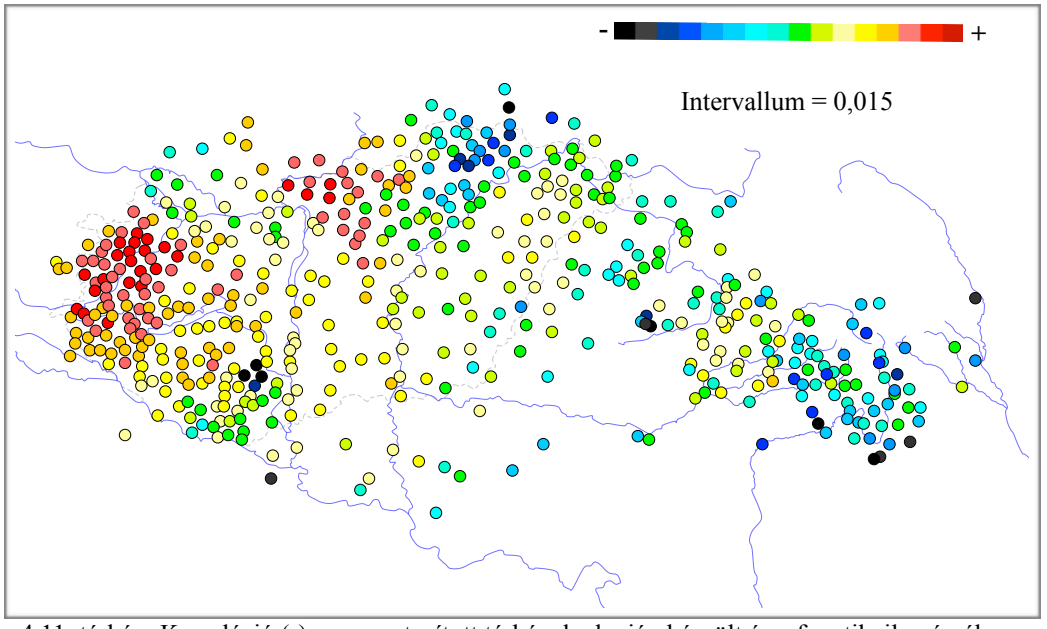
A csoportosítás hangtani, morfológiai (morfofonológiai), illetve lexikai szempontból történhet. A MNyA. és a RMNyA. informatizált korpuszából 127 integrált térképlap alapján, Goebl módszerét követve, összesen 245 munkatérképet hoztunk létre. A munkatérképek közül 197 fonetikai, 16 morfológiai/morfofonológiai, 32 lexikai. Valamennyi kiválasztott térképlap a MNyA. teljes kutatóponthálózatú térképei közül való, amelyeknek többsége leginkább fonetikai szempontok szerint csoportosítható. A térképek kiválasztásakor arra törekedtünk, hogy a magyar nyelvjárások osztályozásakor hagyományosan lényegesnek tartott nyelvi változók (pl. *í~é, ë~ö~e, á~ā, a~à* váltakozás, ikes vs. iktelen ragozás, diftongálás) szerepeljenek elemzési szempontként, lehetőleg hasonló súllyal. Fontos szempont volt továbbá, hogy csak olyan munkatérképeink legyenek, amelyek mutatnak valamilyen jellegzetes területi megoszlást (az elemzési módszerről bővebben l. Kocsis–Vargha 2016).

A kutatópontok közti hasonlóság kiszámításakor, ha az adatok csoportosításából indulunk ki, egy adatpár összevetésekor mindig két lehetőségünk van: vagy azonos csoportba tartoznak (100%-os egyezés), vagy különbözőbe (0%-os egyezés). Egy kutatóponton természetesen több adat is előfordulhat, ekkor a különböző adatösszevetések átlagával számolunk, figyelembe véve az egyezéseket és a különbözőségeket is. Az összes munkatérkép hasonlósági értékeit átlagolva kapjuk meg a kutatópontok közti hasonlóság mértékét, vagyis a települések közti összevetések eredményét megmutató hasonlósági mátrixot.

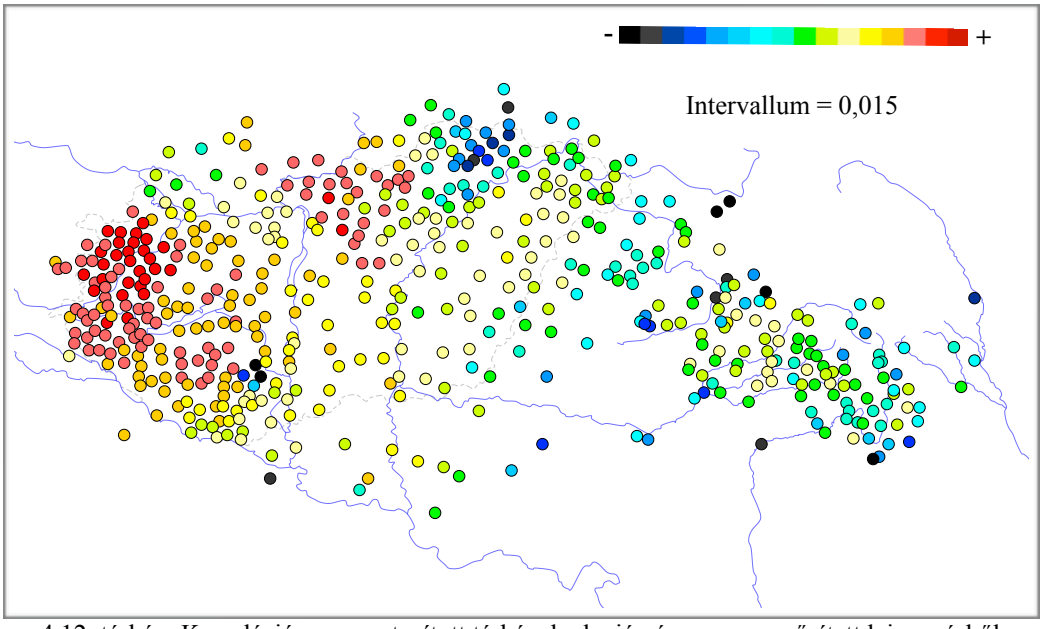
A korábbi fejezetekben már bemutatott, mátrixok közti korrelációs elemzések mintájára összevethetjük egymással az elemzésbe bevont 127 térképlap csoportosításon alapuló nyelvi hasonlósági mátrixát és ugyanazon térképlapok automatikus adatösszevetéssel készült mátrixait. Mantel-teszttel nézve a csoportosítással létrehozott mátrix mind az eredeti, mind az egyszerűsített lejegyzésen alapuló, sőt, a lejegyzés lexikai szintű egyszerűsítésével létrehozott mátrixszal is hasonlóan erősen korrelál. A korrelációs együttható a csoportosítás alapján készült és a finom fonetikai lejegyzésen alapuló mátrix között 0.8902479, a csoportosítás alapú és az egyszerűsített lejegyzés mátrixa között 0.8998361, a csoportosításalapú és a lexikai mátrix között 0.9043933.

A csoportosítás alapján készült és az automatikus adatösszevetéssel létrehozott mátrixok között a Pearson-korreláció mértékét az egyes kutatópontok esetében a 4.11–4.13. térképek mutatják meg. Nagyobb mértékű eltérést (fekete kutatópontok) csak az első két térképen láthatunk, amely az eredeti, mellékjelezett, illetve az egyszerűsített lejegyzés mátrixával való összevetés eredményeit szemlélteti. A legalacsonyabb *r* érték 0,65 (Kéty), amely még mindig jelentős pozitív együttjárásnak tekinthető. A legtöbb azonossághoz közelítő egyezést a második, vagyis az egyszerűsített lejegyzés mátrixával való összevetéskor mértük, erre utal, hogy ezen a térképen van a legtöbb pirosas, narancsos árnyalatú kutatópont. Ugyanakkor a palócföld keleti szélén, illetve attól keletre egy összefüggő sávban, továbbá elszórtan még néhány település esetében jellemzően kisebb az együttjárás a két mátrix között.

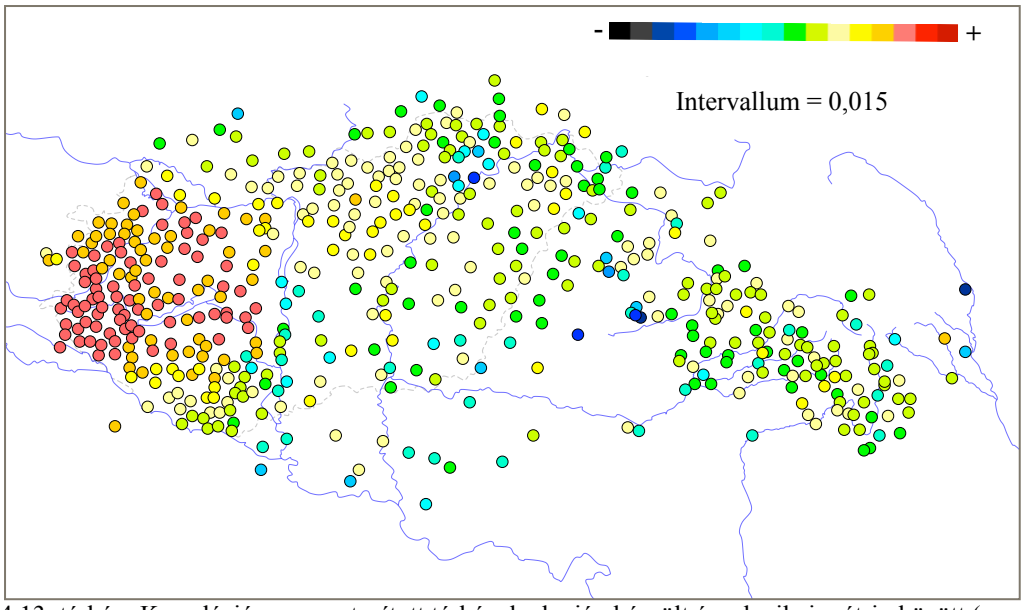
A korrelációs térkép (4.13.) szerint a lexikai jellegű mátrix és a csoportosításon alapuló mátrix minden kutatópont esetében erős együttjárást mutat; a Dunántúlon, különösen nyugaton, az *r* értéke számos ponton közelíti az 1-et, vagyis szinte teljes azonosságra utal. A térkép jól összevethető a 3.3., a MNyA. fonetikailag érzékeny és lexikai jellegű elemzésének korrelációs térképével, ahol a Nyugat-Dunántúl más területekhez képest szintén erősebb együttjárást mutat.



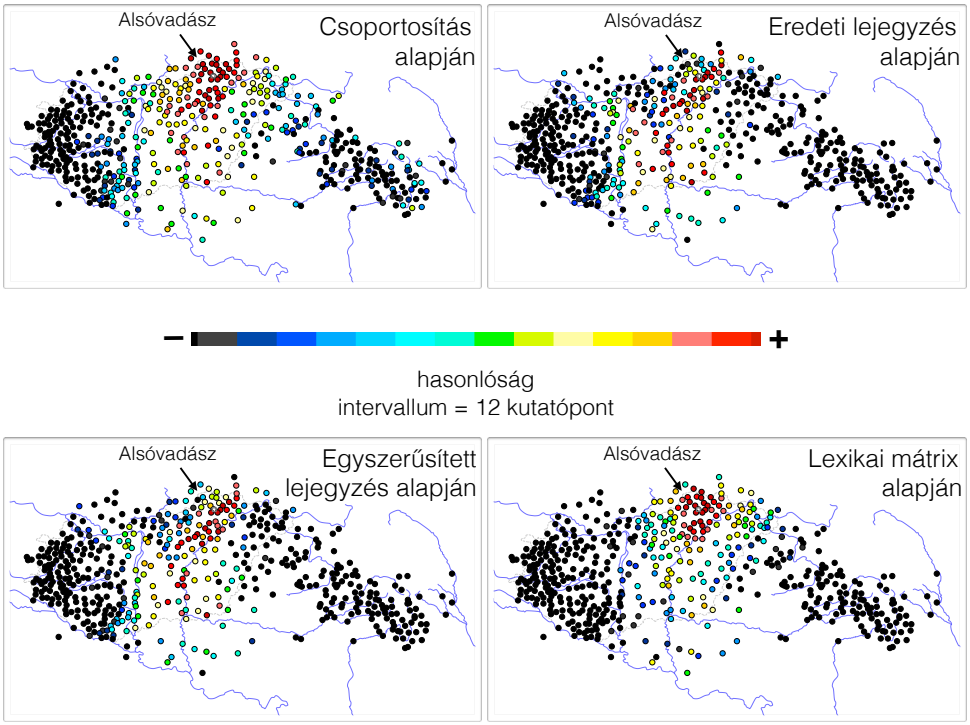
4.11. térkép: Korreláció (r) a csoportosított térképek alapján készült és a fonetikailag érzékeny mátrix között (az *r* értékét a skála szerint a kutatópontok színe mutatja meg, forrás: Kocsis–Vargha 2016)



4.12. térkép: Korreláció a csoportosított térképek alapján és az egyszerűsített lejegyzésből készült mátrix között (az *r* értékét a skála szerint a kutatópontok színe mutatja meg, forrás: Kocsis–Vargha 2016)



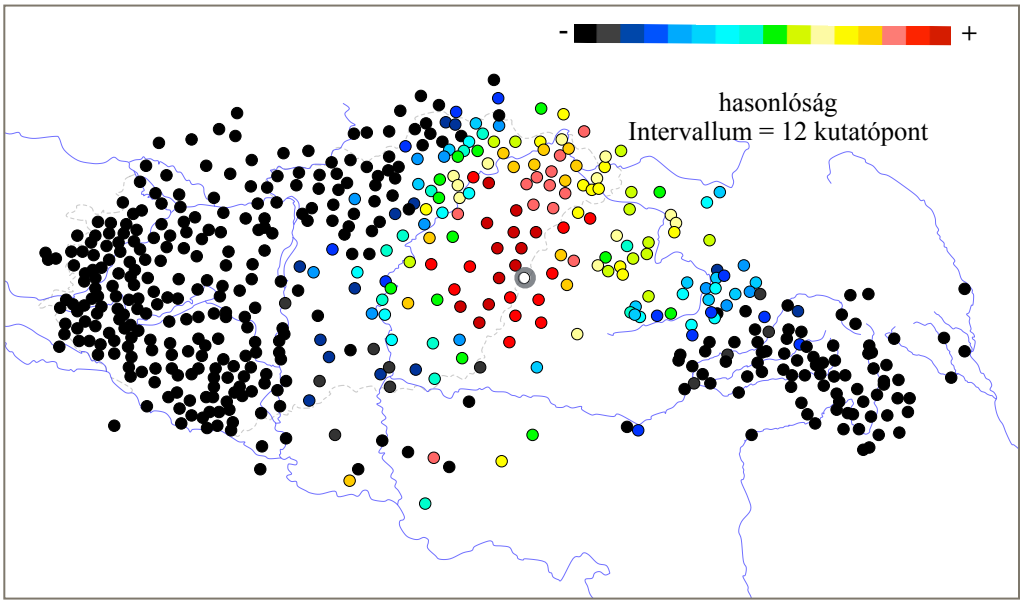
4.13. térkép: Korreláció a csoportosított térképek alapján készült és a lexikai mátrix között (az *r* értékét a skála szerint a kutatópontok színe mutatja meg, forrás: Kocsis–Vargha 2016)



4.2. ábra: Alsóvadász dialektometriai térképei különböző mátrixok alapján

A nyugat-dunántúli kutatópontok nyelvi hasonlósági viszonyai tehát különböző módon elemezve is hasonlóak. Más területeken azonban a különböző nyelvi szintek meghatározó szerepének fölerősítése (a lejegyzés egyszerűsítésén keresztül) eltérő eredményeket hozhat. A palóc nyelvjárásterület keleti peremén, illetve attól keletre elhelyezkedő, a mátrixok között kisebb mértékű korrelációt mutató kutatópontok különböző elemzési módszerrel készült dialektometriai térképei várhatóan számottevően eltérnek egymástól.

Az egyik csekély korrelációt mutató kutatópont Alsóvadász (különösen a 4.12. térképen), a különböző mátrixok szerint kirajzolódó nyelvi hasonlósági viszonyait a 4.2. ábra szemlélteti.[7] A legfeltűnőbb különbség a térképek között, hogy míg a finom fonetikai és a valamelyest egyszerűsített fonetikai lejegyzés alapján készült mátrixok esetében Alsóvadász a palóc nyelvjárásterület határától keletre elhelyezkedő településekkel mutat leginkább hasonlóságot, a tőle nyugatra lévő (palóc jellegű) kutatópontokkal kevésbé, addig a csoportosítással készült és a lexikai jellegű mátrix alapján eltérő mintázat rajzolódik ki. Utóbbi két esetben számottevő hasonlóságot láthatunk Alsóvadász és a tőle nyugatra fekvő, palóc nyelvjárású kutatópontok között.



4.14. térkép: Ártánd dialektometriai térképe a csoportosított térképekből kialakított mátrix alapján

Alsóvadász dialektometriai térképeinek összevetése alapján úgy látszik, a Palócföld határai elég élesen jelen vannak a finom fonetikai és a kisebb mértékben egyszerűsített lejegyzés alapján készült mátrix esetében, a csoportosításon alapuló és a lexikai hangsúlyú mátrix használatakor azonban elmosódnak. A csoportosításon és a lexikai

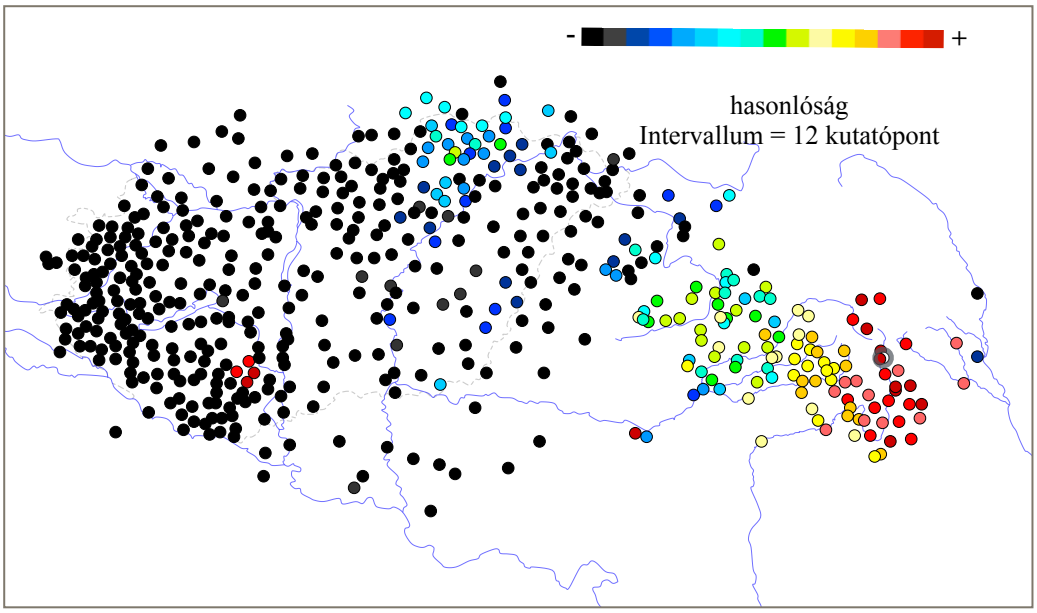
7 A térképek azonos beállítással készültek, a színek 12 kutatópontonként változnak, a legnagyobb mértékű hasonlóságot, akár a korábbi térképeken, a vöröses árnyalat jelzi. A csoportosítás alapján készített mátrix térképén azonban több kutatópont is pirosas színekben látszik, ennek az az oka, hogy sok az azonos mértékben hasonlóságot mutató kutatópont, ilyenkor a beállításban megadottnál több kutatópont is kaphat ugyanolyan színt.

szintű egyszerűsítésen alapuló dialektometriai elemzés azonban mégsem lesz azonos. Bár a csoportosítás alapján készült térképen Alsóvadász olyan kutatópontokkal is mutat nagyobb mértékű hasonlóság, amelyekkel a finom fonetikai és az egyszerűsített lejegyzés alapján nem, az utóbbi két térképen kirajzolódó terület sem vész el a csoportosítás alapú mátrixnál sem. Sőt, a csoportosítás alapú térképen még azok a délebbre fekvő kutatópontok is hasonlónak látszanak, amelyek ott vannak a finom fonetikai és az egyszerűsített lejegyzés alapján készült térképen is. A lexikai jellegű mátrix alapján készült térképen azonban eléggé egyértelműen a földrajzi közelség a meghatározó; a földrajzilag – akár keleti, akár nyugati irányban – közeli kutatópontok mutatnak nagyobb hasonlóságot, és a nyelvi hasonlóság mértéke a földrajzi távolsággal csökken.

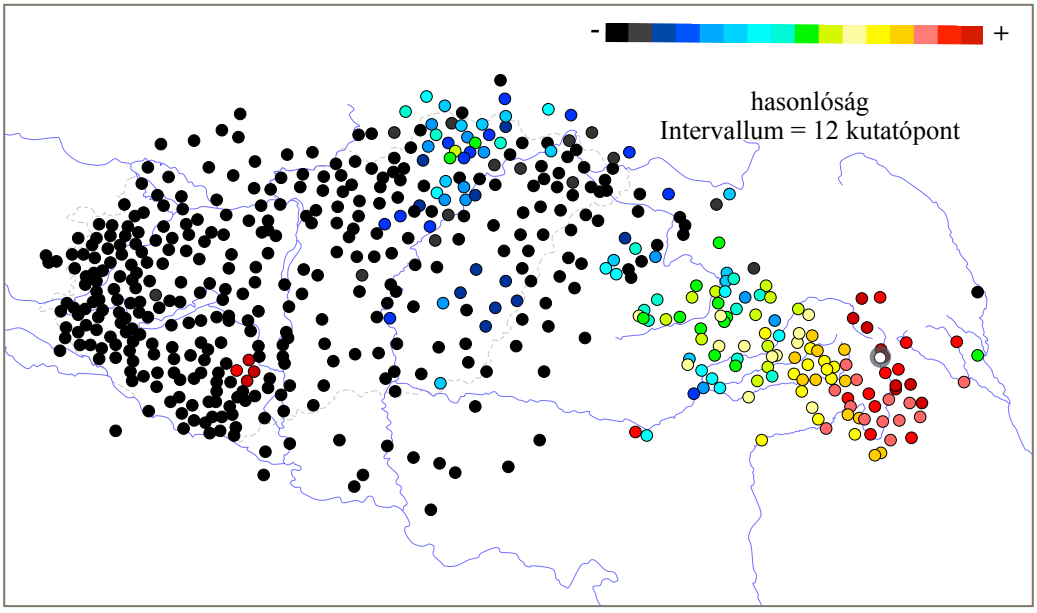
A csoportosítás tehát, bár közel áll hozzá, nem hoz a lexikai jellegű elemzéssel azonos eredményt, hiszen a fonetikai hasonlóságot is bizonyos mértékben tükrözi, ugyanakkor a Palócföld más térképeken éles kontúrjai itt elmosódnak. A fonetikai különbségeket nem megfelelően visszaadó eredmény minden bizonnyal azzal magyarázható, hogy az egyes nyelvi jelenségek nem tényleges (atlaszbeli) előfordulási gyakoriságuknak megfelelően szerepeltek csoportosítási szempontként a munkatérképek létrehozásakor. Így a palóc nyelvjárások legfőbb jellegzetessége – az illabiális *ȧ* és a labiális hosszú *ā* megléte – nem kapott kellő súlyt az elemzésben. A csoportosításkor csupán arra törekedhettünk, hogy lehetőleg minden, a magyar nyelvjárások osztályozásában lényegesnek tekintett változó szerepeljen. Az automatikus elemzéskor azonban az egyes nyelvi jelenségek gyakorisága feltétlenül hatással van az eredményekre, hiszen ott nem kutatói döntések függvénye, mely térképen milyen szempont szerint vetjük össze az adatokat. Vagyis míg az automatikus elemzés során minden különbség, még a legapróbb is, hatással van a két adat közti nyelvi távolság meghatározására, a kutatói csoportosítás során egy munkatérképen csak egyetlen szempont érvényesül.

Érdemes megvizsgálnunk néhány kutatópontot az adattárhatás érvényesülésének szempontjából is, hiszen a csoportosítás egyik nagy előnye lehet az automatikus elemzéssel szemben, hogy itt nem érvényesülnek az adattárak közti, az eltérő lejegyzési szokásokból adódó különbségek. Az előző alfejezetben Ártánd példáján keresztül néztük meg, mi történik a két adattár határán lévő MNyA. kutatópontokkal. (Emlékeztetőül: azért nagyobb kihívás a MNyA. felől közelíteni a kérdéshez, mert ennek az adattárnak sokkal több kutatópontja van, így a lejegyzési stílus azonossága a MNyA. kutatópontok esetében jobban erősíti az adattár kutatópontjai közti hasonlóságot.) A lejegyzés egyszerűsítése valamelyest csökkentette az országhatár nyelvjárási törésvonal jellegét. A csoportosítás alapján készült dialektometiai térkép tanúsága szerint (4.14. térkép) ez a módszer valóban még jobban semlegesítheti az adattárhatást, hiszen az Ártándtól keletre lévő, már a RMNyA.-hoz tartozó kutatópontok nagyobb nyelvi hasonlóságot mutatnak. Ugyanakkor a nyelvi hasonlóság földrajzi súlypontja továbbra is egyértelműen nyugati irányba mutat.

Vacsárcsi és Csíkrákos dialektometriai térképei a csoportosítás alapján kiszámított értékek alapján szinte teljesen azonosak (4.15 és 4.16. térkép). Maguk a kirajzolódó földrajzi hasonlósági viszonyok az egyszerűsített lejegyzés mátrixából készített térképekre hasonlítanak (4.5. és 4.6. térkép), ám azonos beállítással a csoportosítás alapján készült térképeken a dunántúli, Balaton környéki kutatópontok feketék (kevéssé hasonlók).



4.15. térkép: Csíkrákos dialektometriai térképe a csoportosított térképekből kialakított mátrix alapján



4.16. térkép: Vacsárcsi dialektometriai térképe a csoportosított térképekből kialakított mátrix alapján

A csoportosítás alapján készült nyelvi hasonlósági mátrix tehát, bár erősen korrelál az automatikus adatösszevetések eredményeivel, jellegzetesen el is tér azoktól. Jóllehet az elemzés alapját főleg fonetikai szempontok szerint kialakított munkatérképek adják, az egyes jelenségek súlya az elemzésben nem követi azok gyakoriságát. Mint láttuk, ez lehet a magyarázat arra, hogy a Palócföld határa, amelyet a finom fonetikai és az egyszerűsített lejegyzés alapján kialakított mátrixok markánsan megjelenítenek, elmosódik. Elképzelhető, hogy hasonló magyarázatot adhatunk arra is, hogy a fonetikailag érzékeny elemzés alapján nagyobb hasonlóságot mutató távolabbi, dunántúli kutatópontok szintén „eltűnnek" Csíkrákos és Vacsárcsi csoportosítás alapú dialektometriai térképein (l. még a nyelvjárások statisztikai módszerekkel történő, automatikus klasszifikációit bemutató térképeket is az 5.6. alfejezetben).

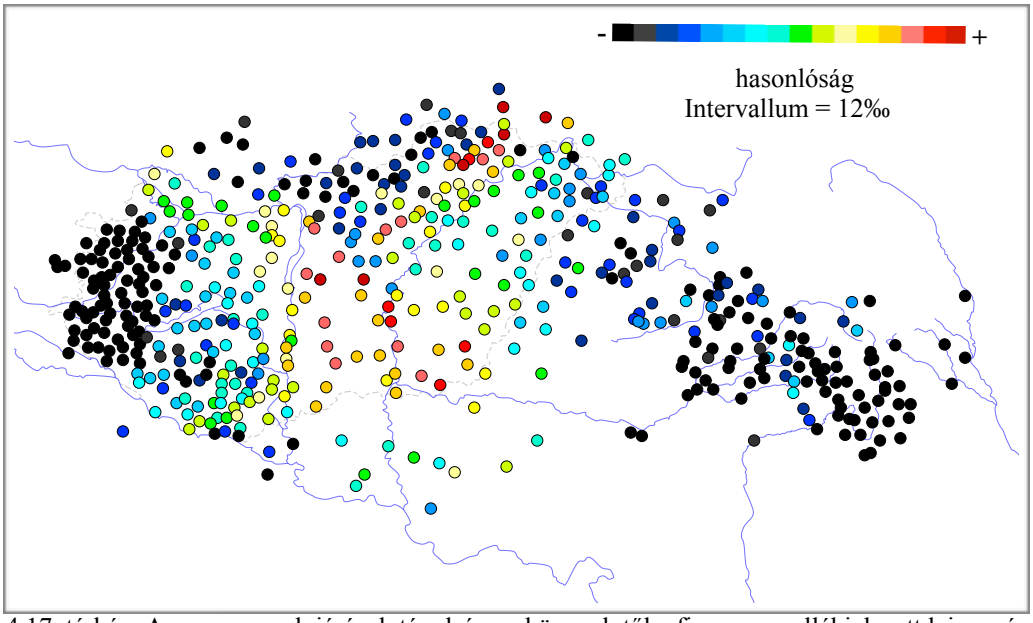
Az adattárhatás egyértelműen a kutatói döntésekre alapuló módszerrel tompítható leginkább, hiszen így a lejegyzési szokásokban rejlő különbségek nem érvényesülhetnek. Elképzelhető, hogy több munkatérkép kialakítása, az egyes változók gyakoriságának tudatos figyelembevétele és érvényesítése az adatok klasszifikációja során kedvezően befolyásolhatja az ezzel a módszerrel megrajzolt nyelvi hasonlósági mintázatok alakulását. Több munkatérkép kialakítása lehetőséget adhat arra is, hogy jelenségcsoportonként külön mátrixokat hozzunk létre, összehasonlítva egymással a különböző változók megrajzolta képet, közelebb kerülve ezáltal az egyes jelenségcsoportok konkrét hatásának kimutatásához az egyes nyelvjárások kapcsolatainak megrajzolásában, nyelvjárási régiókba, csoportokba sorolásukban. A MNyA. és a RMNyA. korpusza alapján aligha lennének vizsgálhatók a szupraszegmentális fonetikai vagy a szintaktikai jelenségek, utóbbiak dialektometriai elemzéséhez Hegedűs Attila kutatásai (2012) szolgáltathatnak alapot.

4.4. A magyar nyelvjárások és a köznyelv

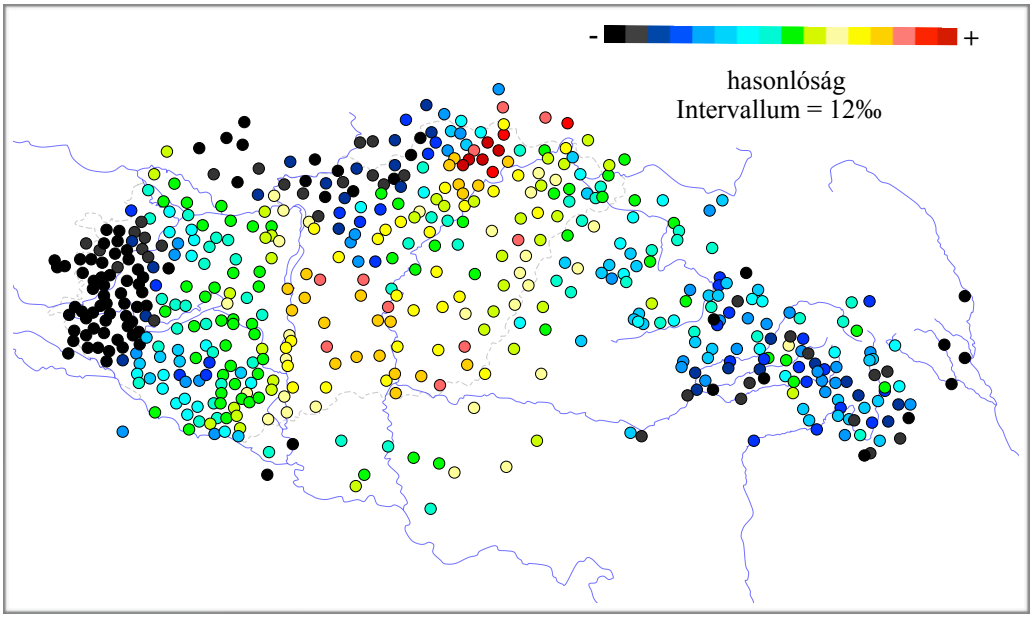
A MNyA. és a RMNyA. integrálása révén a teljes magyar nyelvterületet lefedő dialektometriai térképeket készíthetünk. Ha a kutatópontjainkat kiegészítjük egy fiktív kutatóponttal, amelyen minden térkép esetében a standardnak megfelelő alako(ka)t adjuk meg, akkor elkészíthetjük a kutatópontjaink köznyelvtől való távolságának térképét. A módszert Goebl alkalmazta először a francia nyelvjárások dialektometriai elemzésekor, és később más nyelvatlaszok esetében is (l. Goebl 2002, 2008).

A MNyA. és a RMNyA. integrált, Levenshtein algoritmusával történő elemzése 482 integrált térképlap alapján készült. A „köznyelvi kutatóponton" ezen térképek címszavának korabeli standard változata szerepel. A köznyelvi változatok megadásánál általában a MNyA. térképlapjának címszavát vettem alapul, akkor tértem el attól, ha esetleg valamely változat sokkal gyakoribb volt, mint a címszó (pl. a *burgonya* mellett a *krumpli* is szerepel, illetve az *innám* mellett az *innék* is), illetve ha két hasonló gyakoriságú adat van egy térképen, és egyik változatnak sincs jellemző térbeli eloszlása, ebben az esetben is mindkettőt fölvettem (pl. *bajsza, bajusza*).

A köznyelvtől való távolságról két térképlap is készült. Az egyik az eredeti, finoman mellékjelezett lejegyzésből kiindulva (4.17. térkép), a másik a lejegyzés egyszerűsített, mellékjeleket nem tartalmazó változatának Levenshtein-alapú elemzésével. A MNyA. kutatópontjainak adatai feltehetően finoman mellékjelezett változatukban is jól összevethetők a köznyelvi kutatóponton megadott adatokkal, a RMNyA. esetében azonban, ahol gyakoribb és változatosabb a mellékjelek használata, feltehető, hogy nagyobb nyelvi távolságot mérünk a mellékjelezett, mint az egyszerűsített lejegyzésű változat szerint. A térképek visszaigazolni látszanak ezt a hipotézist.



4.17. térkép: A magyar nyelvjárások távolsága a köznyelvtől a finoman mellékjelezett lejegyzés alapján



4.18. térkép: A magyar nyelvjárások távolsága a köznyelvtől az egyszerűsített, adattárak integrálására alkalmasabb lejegyzés alapján

A térképeken, a korábban látott, egy-egy kutatópont nyelvi hasonlósági viszonyainak ábrázolásához hasonlóan, a köznyelvvel nagyobb hasonlóságot mutató települések meleg, pirosas, narancsos árnyalatot kapnak, a legkevésbé hasonló települések pedig kék vagy fekete színűek. A színek a színskálának megfelelően alakulnak, ám ebben az estben nem meghatározott számú (pl. 10 vagy 12) település esik egy kategóriába, hanem egy szín 1,2%-nyi intervallumnak felel meg. Ez azt jelenti, hogy a köznyelvvel leginkább hasonlóságot mutató településsel kezdődően (ez a 4.17. térképen Nagyszalánc, a hasonlóság mértéke 87,3%; a 4.18. térképen Erdőhorváti, a hasonlóság mértéke 90%) 1,2%-onként haladunk a vöröstől a színskála másik végén lévő fekete felé.

A 4.17. (eredeti lejegyzés szerinti) és a 4.18. (egyszerűsített lejegyzés mátrixa alapján készült) térképen a MNyA. kutatópontjai közül egyaránt a Nyugat-Dunántúl és a Palócföld esik a legtávolabb a köznyelvtől. A RMNyA. kutatópontjainak többsége azonban jelentős különbséget mutat, különösen a székelyföldi kutatópontok, amelyek a 4.17. térképen szinte kizárólag feketében látszanak, a 4.18. térképen azonban már jellemzően nem a legnagyobb távolságot mutató települések közé kerülnek.

A két mátrix alapján a köznyelvvel legnagyobb és legkisebb mértékű hasonlóságot mutató öt-öt települést a 4.2. és a 4.3. táblázat mutatja be. A legnagyobb hasonlóságot mutató kutatópontok pontos sorrendje változik ugyan a mátrix függvényében, de ez a változás abban az értelemben nem jelentős, hogy e települések földrajzilag nagyon közel vannak egymáshoz, valamennyi a történelmi Abaúj és Zemplén megye területén, Sárospatakhoz, a nyelvújítás egykori központjához közel fekszik. Az egyetlen kivétel Gyón, amely az egyre inkább a köznyelv irányában változó Pest megyei települések közé tartozik. Valamennyi kutatópont a MNyA.-ban található.

Mellékjeles adatokból			**Mellékjelek nélkül**		
Kutatópont	**Megye**	**Hasonlóság**	**Kutatópont**	**Megye**	**Hasonlóság**
Nagyszalánc	Abaúj	87,3%	Erdőhorváti	Zemplén	90,0%
Deregnyő	Zemplén	86,9%	Fony	Abaúj	89,8%
Alsódobsza	Zemplén	86,8%	Hernádkércs	Abaúj	89,4%
Nagybózsva	Abaúj	86,4%	Tállya	Zemplén	89,4%
Gyón	Pest	86,4%	Alsódobsza	Zemplén	89,2%

4.2. táblázat: A legnagyobb mértékű hasonlóságot mutató kutatópontok a 4.17. és a 4.18. térképen

A köznyelvtől legtávolabbi kutatópontok nem változnak a mátrix függvényében, mindenképp a négy moldvai RMNyA.-kutatópont és a szerémségi Kórógy adatai mutatnak a legkisebb arányban egyezést a köznyelvi változatokkal. Feltételezhető, hogy ez leginkább ezen kutatópontok hosszú ideje tartó elszigeteltségéből, a köznyelv föléboltozódásának hiányából adódik. Vagyis ezen kutatópontokra tudtak legkevésbé hatással lenni a magyar nyelv központosítására, egységesítésére irányuló törekvések.

Mellékjeles adatokból			Mellékjelek nélkül		
Kutatópont	**Régió**	**Hasonlóság**	**Kutatópont**	**Régió**	**Hasonlóság**
Szabófalva	Moldva	48,8%	Szabófalva	Moldva	54,6%
Bogdánfalva	Moldva	49,7%	Bogdánfalva	Moldva	59,3%
Diószeg	Moldva	52,7%	Diószeg	Moldva	64,2%
Pusztina	Moldva	53,3%	Pusztina	Moldva	66,1%
Kórógy	Szerémség	55,4%	Kórógy	Szerémség	66,2%

4.3. táblázat: A legkisebb mértékű hasonlóságot mutató kutatópontok a 4.17. és a 4.18. térképen

A hagyományosan is megkülönböztetett nyelvjárási régiók közül a nyugat-dunántúli és a palóc régió amellett, hogy jelentős távolságot mutat a köznyelvtől, markánsan elkülönül a nyelvterület többi részétől is (e vonatkozásban lásd az 5.6. fejezetben az 5.14. ábrán a fonetikailag többé-kevésbé érzékeny mátrixok négy kutatópont-csoportot elkülönítő klasztertérképeit).

Kínálkozik a felvetés, vajon hatással van-e a köznyelvtől (illetve a többi nyelvjárástól) való távolság a nyelvjárások megítélésére. Elsőként Imre Samu tanulmányát érdemes említenünk (1963), amelynek eredeti célja az volt, hogy megállapítsa, hol beszélnek a legszebben magyarul. Kutatását azonban nem objektíven elemzett beszélői vélekedésekre, hanem egyrészt a nyelvjárásgyűjtés során szerzett tapasztalataira, benyomásaira, másrészt saját értékítéletére alapozta. Ez utóbbi leginkább abban állt, hogy a magyar nyelvjárások jellemzésére leginkább használt nyelvi változókat sorra véve mindig a köznyelvi (budapesti) variánsnak megfelelőt ítélte szebbnek, így valójában kis túlzással azt mondhatjuk, fejben készített a fentebbihez hasonló dialektometriai elemzést a MNyA. gyűjtőmunkálataira, adataira alapozva. Nem csoda hát, ha a fenti térképekre szépen rímelő végkövetkeztetésre jut: „Mindezek alapján én a végső helyezésben a fenti sorrendet tartanám meg, tehát 1. Sárospatak környéke, 2. Budapest, 3. Budapest környéke."

A kérdést objektív módszerekkel kívánta vizsgálni a Magyar Nemzeti Szociolingvisztikai Vizsgálat (a továbbiakban MNSZV), ahol megkérdezték az adatközlőket arról, hogy az ország különböző pontjai közül hol beszélnek a legszebben és hol a legcsúnyábban.[8] Az eredményeket Kontra Miklós összegzi (2003), értékelése szerint Budapesten beszélnek a legszebben, illetve az adatközlőkre általában jellemző, hogy saját nyelvváltozatukat tartják a legszebbnek. Kivétel a Nyugat-Dunántúl, ahol erősnek mutatkozik az önstigmatizáció, itt az adatközlők közül legtöbben a Vas megyei beszédet ítélték legcsúnyábbnak.

A MNSZV kutatóponthálózata ugyan csak részben egyezik a MNyA. kutatópont-hálózatával, az egyes MNSZV pontokhoz legközelebbi kutatópontok értékeit alapul véve azonban megvizsgálhatjuk, milyen összefüggés van a köznyelvtől való távolság és a saját nyelvváltozat megítélése között. Az eredmények alapján statisztikailag is szignifikáns mértékben ítélik inkább szépnek saját nyelvváltozatukat azok, akik a köznyelvvel nagyobb mértékű egyezést mutató nyelvjárást beszélnek, mint azok, akik a

[8] Számunkra annyiban szerencsétlen a kérdésföltevés, hogy a határon túli területeket eleve kizárja a lehetséges választás köréből.

köznyelvitől távolibbat. Szintén statisztikailag szignifikánsan nyilatkoznak negatívabban saját nyelvváltozatukról ott, ahol a legnagyobb a köznyelvtől való távolság, jellemzően a Nyugat-Dunántúlon (Vargha 2016a), mint azokon a településeken, ahol nagyobb mértékű a köznyelvvel való hasonlóság. Szombathelyen például hét adatközlő közül senki sem tartotta a legszebbnek saját városának vagy megyéjének nyelvhasználatát, hárman Budapestet, egy-egy adatközlő Pest megyét, Debrecent, illetve az Alföldet nevezte meg. Arra a kérdésre viszont, hol beszélnek a legcsúnyábban, hárman is Vas megyét választották, egy adatközlő pedig a szomszédos Zala megyét, egy-egy személy Szegedet, Szabolcsot és Baranyát.

Az újabban végzett kutatások tanúsága szerint is inkább nyilatkoznak negatívan saját nyelvjárásukról a nyugat-dunántúliak (l. Iglai 2017: 115–160), de olyan területeken is megjelenhetnek a saját nyelvváltozattal kapcsolatos negatív attitűdök, ahol ez a korábbi vizsgálat tanúsága szerint még nem volt általánosan jellemző (Kontra–Németh–Sinkovics 2016: 147–172).

4.5. Összefoglalás

A 4. fejezetben két legnagyobb atlaszunk, a MNyA. és a RMNyA. integrált dialektometriai elemzésével, annak módszertani kérdéseivel és kihívásaival ismerkedtünk meg, az adatok csoportosításán alapuló salzburgi és az automatikus adatelemzést használó groningeni eljárás alkalmazásának tanulságain keresztül. Az atlaszintegrálás legnagyobb problémája, hogy bár mindkét adattár a magyar egyezményes lejegyzést használja, mégis vannak jellegzetes különbségek az adatrögzítés gyakorlatában. A lejegyzési szokásokban megmutatkozó különbségek pedig az adatok automatikus összevetése esetén olyan helyeken is eredményezhetnek nyelvi különbséget, ahol valójában nincsen, ennek hatása (az adattárhatás) aztán természetesen a mátrixban is jelentkezik. A problémára többféle megoldás is kínálkozik, ilyen lehet a kutatói csoportosításon alapuló adatelemzés, illetve az adatok lejegyzési pontosságának, fonetikai információgazdagságának az integrálás szempontjából optimális egyszerűsítése. A csoportosításon alapuló elemzés valóban igen hatékony az adattárhatás tompításában, az egyes, a nyelvjárások osztályozását meghatározó nyelvi jelenségeket azonban – a csoportosításkor meghatározott szempontok miatt – nem azok előfordulási aránya szerint veszi figyelembe. Az adatok automatikus egyszerűsítése valamivel kevésbé hatékonyan, de szintén jól alkalmazható az adattárhatás tompítására. Megjegyzendő azonban, hogy bár a kutatópontok többségénél az egyszerűsítés a nyelvi hasonlósági mintázatokon nem változtat, nyelv(járás)szigetek esetében a térben és időben távoli nyelvi kapcsolatok kimutatására kevésbé jól használható, mint az eredeti, finoman mellékjelezett lejegyzés.

A MNyA. ás a RMNyA. integrált korpuszát kiegészíthetjük egy fiktív köznyelvi kutatóponttal, amelyhez minden címszó esetében megadjuk a standardnak megfelelő adato(ka)t. A köznyelvi kutatópontot is tartalmazó mátrix alapján mérhetővé és térképezhetővé válik az egyes nyelvjárások köznyelvtől való távolsága. Az elemzések alapján a köznyelvtől (és a nyelvterület többi részétől is) a palóc és nyugat-dunántúli nyelvjárások különülnek el a leginkább (nem számítva a köznyelvi föléboltozódás hiányával jellemezhető, nyelvjárásszíget helyzetű kutatópontokat). Összefüggés mutatkozik a nyelvjárás-köznyelv távolság és a beszélők saját nyelvváltozatukról alkotott esztétikai ítélete között: minél távolabb van egy nyelvváltozat a köznyelvtől, annál hajlamosabbak azt beszélői negatívan megítélni, vagyis más nyelvjárásoknál csúnyábbnak gondolni.

5. Nyelvjárások felosztása, osztályozása dialektometriai alapon

Az 5. fejezet témája a nyelvi hasonlóság térbeli összefüggéseinek feltárása, megmutatása egy nyelvatlasz (vagy akár a teljes nyelvterület) egészén. A következő kérdésekre keressük a választ: Hogyan változnak a statisztikai elemzés eredményeképpen kirajzolódó nyelvjárási területek az alkalmazott mátrix függvényében? Hogyan változnak a nyelvi hasonlóság földrajzi mintázatai egyes területeken annak függvényében, mely adattárak (kisebb, illetve nagyobb területet lefedő, ritkább vagy sűrűbb kutatóponthálózatú atlasz) alapján készül az elemzés? Mi következik ebből a nyelvjárások felosztása, nyelvjárásterületek kialakítása szempontjából?

5.1. Klaszteranalízis és többdimenziós skálázás

A dialektometriai kutatásoknak fontos célkitűzése már a kezdetektől, hogy a kutatópontok közti nyelvi hasonlóság mértéke alapján objektív módszerekkel állapítsanak meg nyelvjáráshatárokat, jelöljenek ki nyelvjárásterületeket. A két leginkább alkalmazott statisztikai eljárás, a klaszteranalízis és a többdimenziós skálázás felhasználását a dialektometriai kutatásokban általánosságban az 1.5. fejezet ismerteti. Itt a fejezetben látható térképek kialakításának folyamatát ismertetem röviden.

A nyelvi hasonlóság kiszámítása minden esetben a Bihalbocs programmal történt, hiszen ez az egyetlen, a magyar egyezményes hangjelölést támogató számítógépes alkalmazás. Ez az eszköz teszi lehetővé - egyedülálló módon - többek között azt is, hogy a fonetikai finomság, információgazdagság különböző szintjein végezhessünk elemzéseket.

A hasonlósági mátrix átalakítását a nyelvi távolságot megmutató mátrixszá, illetve a statisztikai elemzéseket a szabadon letölthető R statisztikai programcsomaggal készítettem. A térképek hátterében lévő Ward-féle klaszteranalízis kiszámításához a *ward.D2* algoritmust használtam. A többdimenziós skálázást szintén R-ben, a *cmdscale* funkció segítségével végeztem el.

A statisztikai elemzések eredményének importálására és térképezésére a Bihalbocsban került sor, megfelelő algoritmusok kidolgozásával és alkalmazásával. A klaszteranalízissel készült térképeknél egyszerűen azonos színeket rendelünk az egy csoportba sorolt kutatópontokhoz. Annak érdekében, hogy a térképek harmonizáljanak egymással, s így összevethetőbbek legyenek, a csoportok színét esetenként utólag képszerkesztő programban módosítottam.

A többdimenziós skálázás (multidimensional scaling, MDS) során három dimenzióval dolgoztam, minden egyes dimenziót egy RGB színkomponensnek feleltetve meg. Így az első dimenzió esetében például, a szélső értékeket alapul véve, a kutatópontok az R-ben kiszámított, és a Bihalbocsba importált érték függvényében kapnak gyöngébb vagy erősebb pirosas színezetet. A kutatópontok színe a három dimenzió mentén a piros, a zöld és a kék szín intenzitásának függvénye.

A statisztikai módszerek alkalmazása a nyelvjárások felosztásában objektívnek mondható, hiszen nem néhány kiragadott példa alapján, esetleges előzetes elvárásokat érvényesítve csoportosítjuk az egyes kutatópontokat, hanem valamely matematikai eljárás alkalmazásával, automatikusan alakul a térképeken a kutatópontok színe, színárnyalata, a hasonlósági mátrix alapján, az alkalmazott eljárás függvényében. Fontos azonban megjegyeznünk, hogy a választott módszer – ha statisztikai eljárások alkalmazásával is – egy lehetséges interpretációját hozza létre az adatok összevetésén alapuló hasonlósági mátrixnak. Más eljárások alkalmazása más térképek létrehozásához vezet(het)ne. Azért döntöttem a Ward-féle klaszteranalízis és a többdimen-

ziós skálázás színárnyalatos térképezése mellett, mert jelenleg ezek a legelfogadottabb, legáltalánosabban használt automatikus csoportosítási és a nyelvjárások közti kapcsolathálózatot vizualizáló technikák a dialektometriában (vö. Goebl 2002, Heeringa 2004, Prokic–Nerbonne 2008, Grieve 2014). A klasztertérképen a kutatópontok kategorikusan elkülönülő csoportokat alkotnak, a hagyományos módszerekkel kijelölt nyelvjárásterületekhez hasonlóan, míg az MDS-térképen megjelenő színárnyalatok a nyelvi kontinuumot mutatják meg.

A következőkben a korábbi fejezetekben már megismert adattárak hasonlósági mátrixaiból létrehozott klaszter- és MDS-térképeken mutatom be a nyelvjárások automatikus felosztásának lehetőségét. Mint ahogyan azt a korábbi elemzésekben láttuk, ugyanazokból az adatokból különböző dialektometriai elemzési módszerekkel (csoportosítás vagy automatikus összevetés, l. 4.3. fejezet), a lejegyzés finomságának, fonetikai információtartalmának függvényében (l. 2.3. és 3. fejezet) különböző mátrixokat hozhatunk létre. Lehetőségünk van továbbá arra is, hogy ugyanazon kutatópontok nyelvi hasonlósági viszonyait a nyelvjárási térképekből valamely szempont szerint kialakított részkorpuszok mátrixainak összevetésével vizsgáljuk (l. Bodó–Vargha 2016). Különböző mátrixok szerint nézve egy-egy kutatópont dialektometriai térképei akár jelentős eltérést is mutathatnak (lásd pl. Csíkrákos, 3.4. fejezet, 3.4. ábra). Az egyes mátrixokat további különféle elemzéseknek alávetve hozhatjuk létre a kutatópontok területi felosztását, illetve a nyelvi kontinuumot bemutató térképeket. A nyelvi hasonlóság földrajzi mintázatainak vizualizálására számtalan lehetőségünk van tehát. Az itt megjelenített térképek éppen ezért a teljesség igénye nélkül, de bőséges példaanyagon mutatják meg a nyelvjárások dialektometriai alapú felosztásának lehetőségét, egyszersmind rávilágítva arra, hogy az objektív, statisztikai módszerek alkalmazása mellett is bizonyos mértékben relatív minden kategorikus csoportosítás, amely az elemzés beállításainak is függvénye. A kategorikusan csoportosító módszerrel szemben ezért a nyelvi kontinuumot feltáró elemzéseket preferálhatjuk.

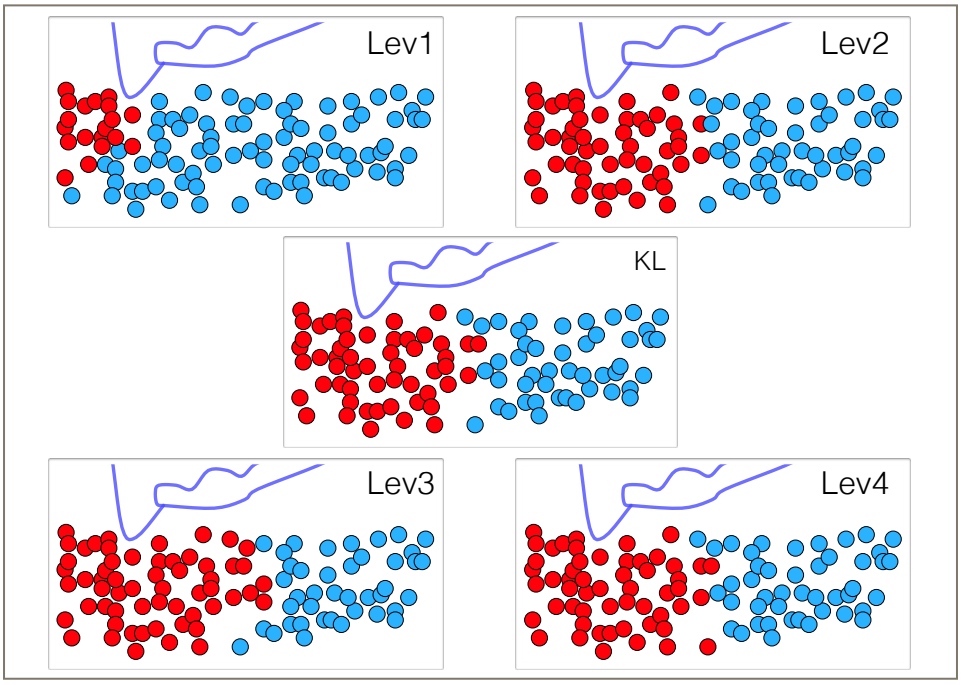
Az aggregált elemzési módszerből alapvetően következik, de nem árt leszögeznünk, hogy a mátrix alapú, statisztikai módszerekkel történő területi felosztás élesen különbözik a hagyományostól. A nyelvjárások elhatárolását célszó, jelenséghatárok kijelölésén alapuló izoglossza-módszerrel a kutató az általa kiválasztott térképlapokon egyes változatok elterjedtségének határait megrajzolva, a jelenséghatárok alkotta izoglossza-nyalábok elhelyezkedéséből kiindulva próbál elkülöníteni nyelvjárásterületeket. A dialektometriai elemzésben az egyes nyelvi jelenségek hatása nem különül el, egyszerre játszik szerepet, az automatizált eljárások esetében az adatokban való előfordulás gyakoriságának függvényében (l. erről még Kocsis–Vargha 2016, illetve a 4.3. alfejezetet).

5.2. A Somogy-zalai nyelvatlasz

A 2. fejezetben öt hasonlósági mátrix mentén vizsgáltuk a S–ZA. kutatópontjai közötti nyelvi hasonlóságot, Mantel-tesztek segítségével összevetve egymással a különböző mátrixokat, illetve a kutatópontonkénti korreláció mértékét térképeken is szemléltetve. A mátrixok közül négyet Levenshtein-algoritmus használatával, egyet Király Lajosnak, a S–ZA.-ban közölt térképenkénti csoportosításai alapján hoztunk létre. Itt az egyes mátrixok klaszteranalízisének, illetve háromdimenziós skálázásának eredményét nézzük meg térképekre vetítve.

A Ward-féle klaszteranalízis eredményét klasztertérképek szemléltetik oly módon, hogy fokozatosan egyre több csoportot különítünk el és jelenítünk meg különböző színekkel a térképeken. Először (5.1. ábra) két csoportra osztjuk a területet. Az első

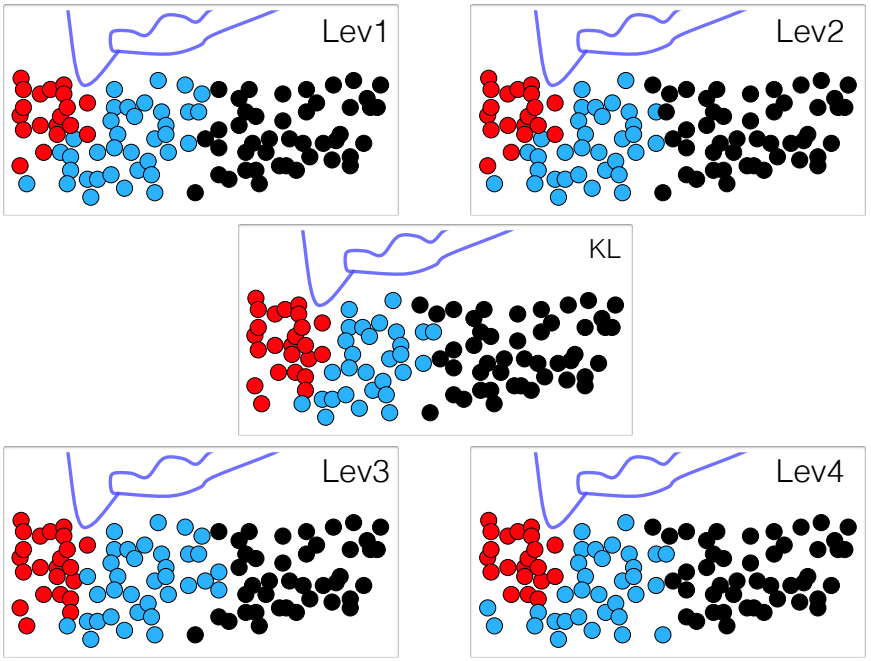
(Lev1), a részletes fonetikus lejegyzés alapján készült elemzés jól elkülönül a másik négytől, a két szín nagyjából az 1950 előtti megyehatár mentén válik szét, míg a másik négy mátrix esetében a gyűjtőterület közepe a határ.



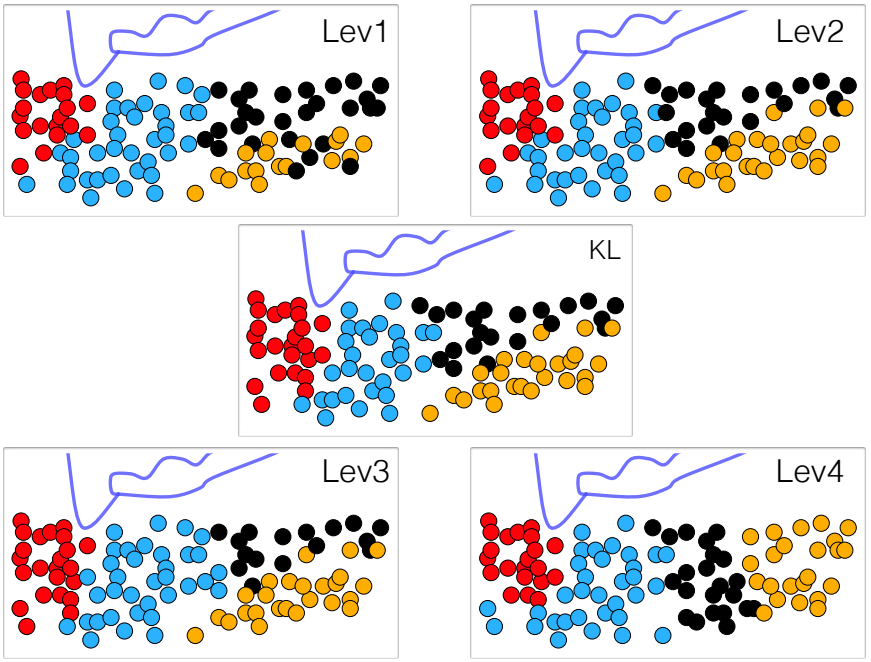
5.1. ábra: A S–ZA. különböző mátrixai alapján készített klaszteranalízisek térképezése a kutatópontok két csoportra osztása esetén

A kutatópontokat három csoportra osztva (5.2. ábra) azt tapasztaljuk, hogy a fonetikai részleteket is tartalmazó lejegyzés mátrixa (Lev1) esetében a somogyi terület középtájt ketté válik, a mellékjeleket nem tartalmazó lejegyzés mátrixa (Lev2) esetében pedig az atlasz gyűjtőterületének nyugati fele válik ketté, a Lev1 mátrixhoz igen hasonló hármas felosztást hozva létre. A lejegyzés további egyszerűsítésével készült Lev3 mátrix és a lexikai szintű elemzésnek megfeleltethető Lev4 mátrix klasztertérképe is csak néhány település esetében tér el, a kutatói csoportosítás alapján készült KL mátrix a nyugati részen a Lev3, a keleti részen inkább a Lev4 mátrixszal mutat nagyobb hasonlóságot.

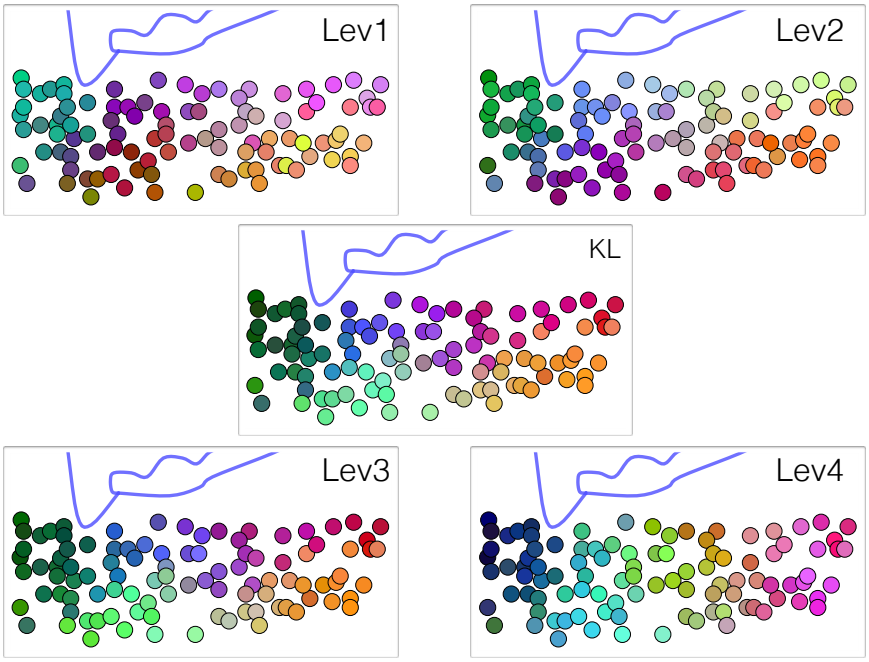
Négy csoport kijelölése (5.3. ábra) a Lev1 mátrix esetében már nem hoz egyértelmű eredményt, az atlasz keleti részén bomlanak két csoportra a kutatópontok, de nem hoznak létre jól körülhatárolható területi egységeket. A Lev2, Lev3 és a KL mátrix klasztertérképe jellegében hasonló a Lev1 mátrixéhoz, a keleti terület észak-déli megoszlást mutat, itt azonban egyértelműbb a két csoport földrajzi elhatárolhatósága. A lexikai különbségekre érzékeny Lev4 mátrix esetében a csoportok közti elkülönülés kelet-nyugati irányú, tehát a földrajzi távolság mentén alakul.



5.2. ábra: A S–ZA. különböző mátrixai alapján készített klaszteranalízisek térképezése a kutatópontok három csoportra osztása esetén



5.3. ábra: A S–ZA. különböző mátrixai alapján készített klaszteranalízisek térképezése a kutatópontok négy csoportra osztása esetén



5.4. ábra: Többdimenziós skálázással készült térképek a S–ZA. mátrixai alapján

Ugyan az egyes nyelvi jelenségek már nem megkülönböztethetők a hasonlósági mátrixok alapján, a hátterükben lévő automatikus egyszerűsítés alapján kikövetkeztethető, hogy az *ö*-zés lehet az a nagyobb súlyú nyelvi tényező, amely a Lev3 mátrix kialakítására még hatással van, de a Lev4 mátrixra már nincsen. Ez lehet a magyarázat arra, hogy a Lev4 mátrix klasztertérképe jellegében lényegesen eltér a többitől.

A klasztertérképeket is könnyebben tudjuk értelmezni, ha vetünk egy pillantást a mátrixok alapján készült MDS-térképekre (5.4. ábra). A kutatópontok itt három dimenzió mentén kapnak színt, így értelemszerűen nem jönnek létre csoportok, illetve azokat elhatároló vonalak. A három összetevőből kikevert színek a nyelvi kontinuumot tükrözik (a színárnyalatok csak térképen belül vethetők össze, térképek között nem, mert a statisztikai eljárás során mindig másként alakul a kutatópontok három dimenzióba rendezése). Összevetve e kétféle térképes elemzést, a Lev1 mátrix esetében jól látszik, hogy az MDS-térképen Zala és Somogy határsávjában élesebben elkülönülnek a színek, mint a többi mátrix MDS-térképein, ami jól egybevág a kutatópontokat két csoportra osztó klasztertérképek kapcsán tapasztaltakkal. Ezen az MDS-térképen a színárnyalatok további, egyértelműnek tűnő felosztást nem tesznek lehetővé.

Az MDS-térképek közül leginkább a Lev2, a mellékjeleket nem tartalmazó, egyszerűsített lejegyzésen alapuló mátrixból készült elemzés alkalmas arra, hogy kisebb csoportokra osszuk a kutatópontokat. Megfigyelhető a színek szétválása Somogy és Zala határán, de a somogyi részeken kelet-nyugati és észak-déli irányú elkülönülés is kivehető. A KL és a Lev3 mátrix alapján készített MDS-térképek szinte teljesen egyeznek, összhangban a mátrixok közti igen erős korrelációval ($r = 0{,}969$), és jellegükben a Lev2 mátrix MDS-térképére hasonlítanak. Az MDS-térképek előnye a kategorikus felosztású klasztertérképekhez képest, hogy a színárnyalatoknak köszön-

hetően egyszerre érvényesül a különböző nyelvi tényezők hatását tükröző észak-déli és kelet-nyugati tagolódás (vö. Király 1990, 2005).

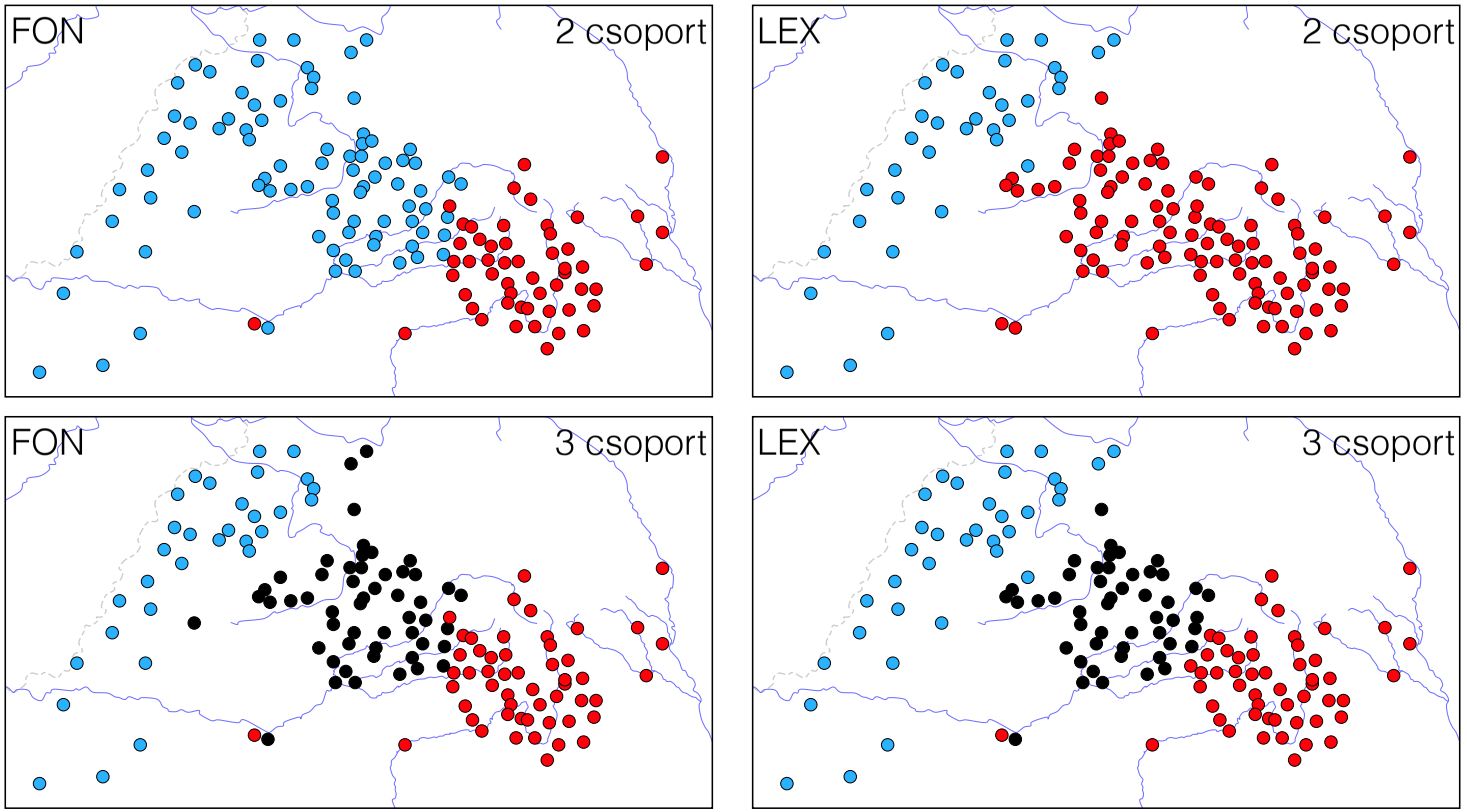
A Lev4 mátrix esetében szintén jó összevetés kínálkozik a klasztertérképek és az MDS-térkép között. A többi mátrixszal ellentétben, itt nem érzékelhető a színárnyalatok alapján a kutatópontok észak-déli megosztottsága, a pontok kelet-nyugati irányú kontinuum mentén rendeződnek el, éles váltás a színek alapján sehol sem érzékelhető, az árnyalatok leginkább a kutatópontok közti földrajzi távolság függvényében változnak. Így a Lev4 mátrixból készített elemzés hasonlóan tér el a többitől, mint ahogyan azt a klasztertérképeknél is láthattuk.

5.3. A romániai magyar nyelvjárások atlasza

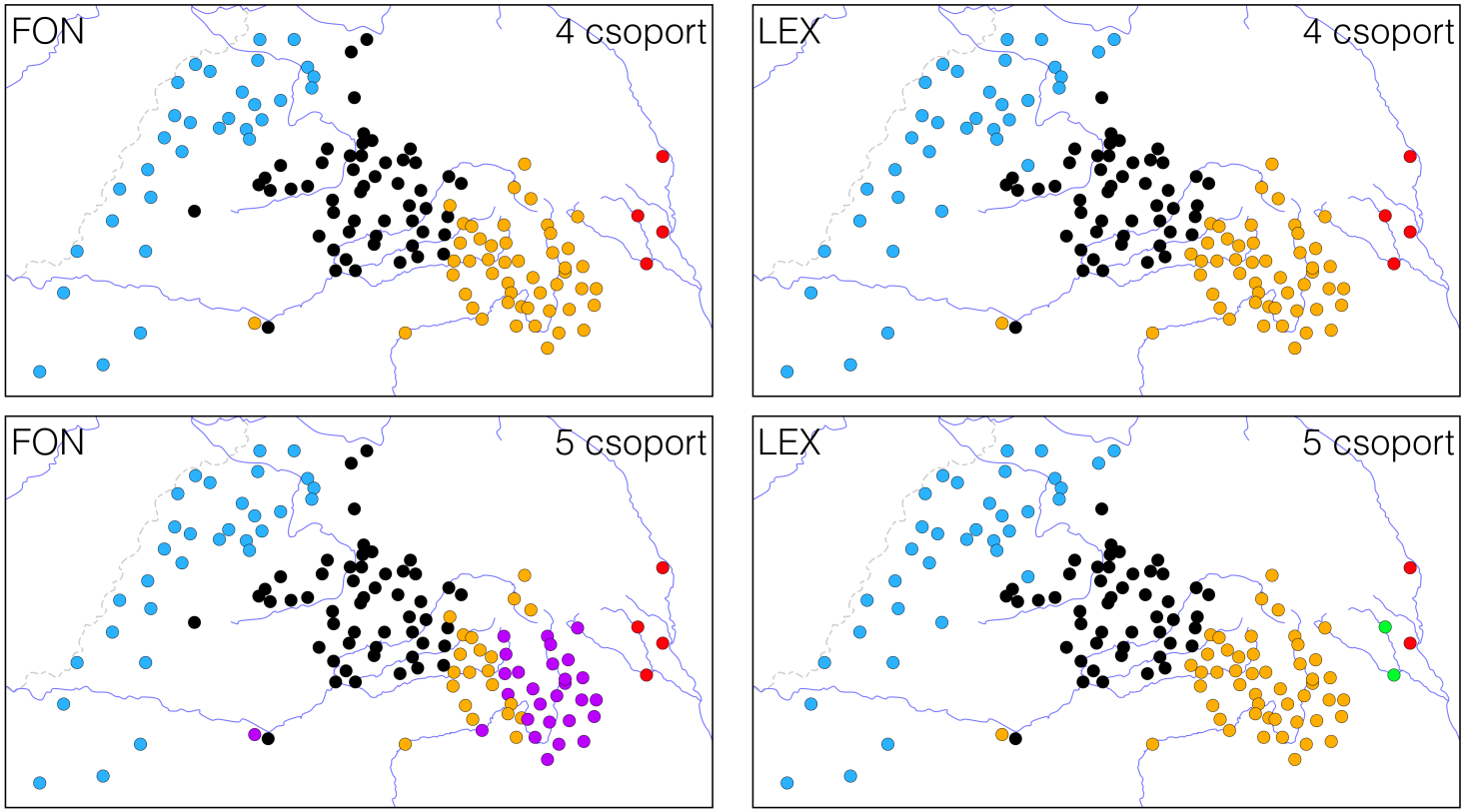
A RMNyA.-ból összesen 1517 térképlapnyi anyag állt rendelkezésre informatizált és ellenőrzött formában az itt bemutatott térképek elkészítésekor, ez valamivel kevesebb, mint a teljes adattár fele, de mindenképp példátlanul nagy adatmennyiség a dialektometriai elemzés szempontjából. (Az automatikus összevetéseken alapuló dialektometriai elemzésekben általánosan felhasznált térképszám általában nem több néhány száz térképlapnyinál, vö. Heeringa 2004.) Az adattár külön előnye, hogy a gyűjtést egyetlen kutató, Murádin László végezte, így nem kell számolnunk a több terepmunkás különböző lejegyzési szokásaiból eredő esetleges inkoherenciával.

A korábbi fejezetekben önállóan nem, csak a MNyA.-val integrálva mutattam be dialektometriai elemzést a RMNyA. adataiból (az atlasz egyes kutatópontjainak nyelvi hasonlósági mintázataira fókuszáló önálló elemzésre l. Vargha 2016b). Itt a lejegyzés eredeti, finoman mellékjelezett változatából készült mátrix és a lejegyzés fonetikai információk szempontjából radikálisan egyszerűsített változatából készült, a lexikai különbségekre érzékeny mátrix klaszter- és MDS-térképeit mutatom be. Mivel mindkét elemzés ugyanazon a korpuszon készült, a fonetikailag érzékeny mátrix kialakításában a lexikai különbségek is szerepet játszanak: akkor beszélhetünk nagy hasonlóságról, ha mind lexikai, mind fonetikai szempontból sok az azonosság az adatok között. A lexikai hangsúlyú mátrix esetében a mellékjelek, magánhangzó-minőségek, sőt, a rokon mássalhangzók közti különbségek sem számítanak, így itt az adatok közti nagyobb, jellemzően lexikai eltérések lesznek csak hatással a nyelvi hasonlóság mértékének alakulására. (A morfológiai eltérések a kvantitatív elemzésben lényegében súlytalanok a vonatkozó atlaszadatok viszonylag kis száma miatt.)

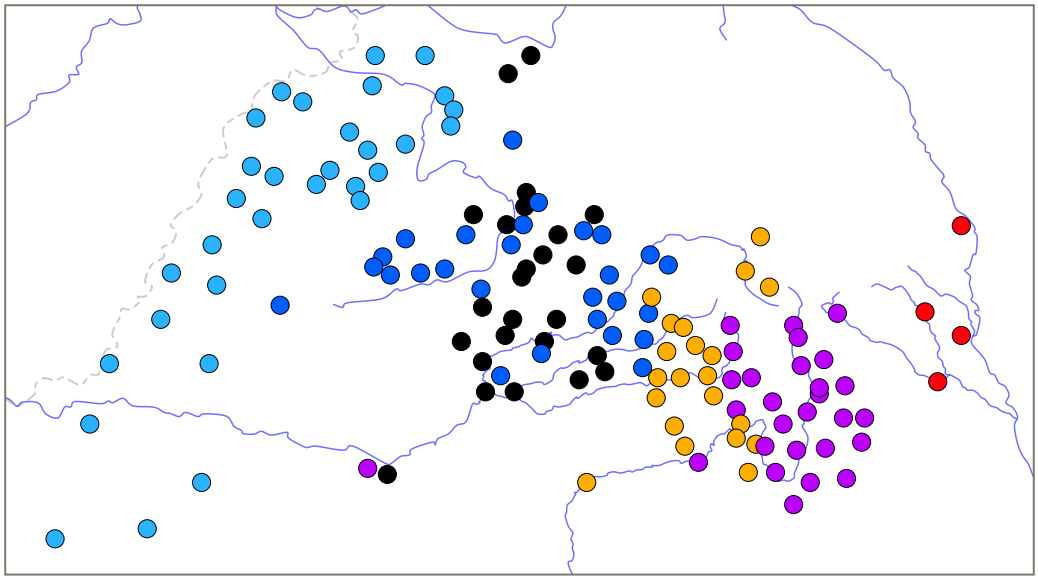
A kutatópontokat két és három csoportra osztó Ward-féle klaszteranalízis alapján készült térképek az 5.5. ábrán láthatók. Két csoport elkülönítése esetén a fonetikai mátrix alapján nagyjából a terület középvonalában, Mezőség és Székelyföld határához közel válnak ketté a kutatópontok, a lexikai mátrix alapján viszont inkább Partium és Erdély megoszlást láthatunk. A kutatópontok három csoportba sorolása esetén a fonetikailag érzékeny mátrix esetében a nyugati terület bomlik ketté nagy vonalakban Partium–Mezőség megoszlást mutatva, a lexikai hangsúlyú mátrix esetében pedig a keleti rész bomlik ketté, a két térkép így szinte azonos. Négy csoport elkülönítése (5.6. ábra) is igen hasonló felosztást eredményez mindkét mátrix esetében, a négy moldvai kutatópont különül el a Székelyföldtől.



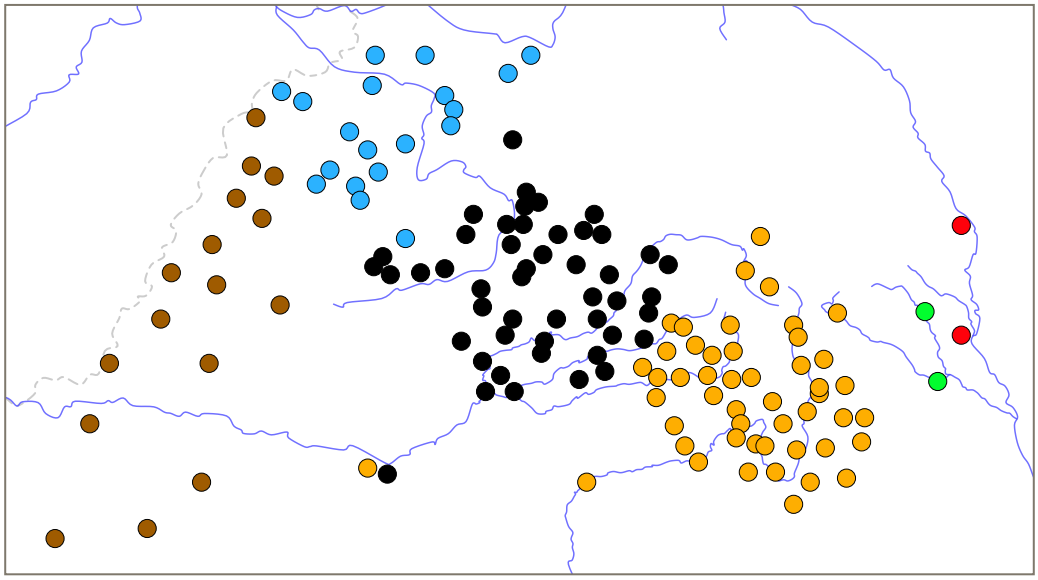
5.5. ábra: A RMNyA. Ward-féle klaszteranalízises térképei különböző mátrixok alapján, 2 és 3 csoport térképezése esetén



5.6. ábra: A RMNyA. Ward-féle klaszteranalízises térképei különböző mátrixok alapján, 4 és 5 csoport térképezése esetén



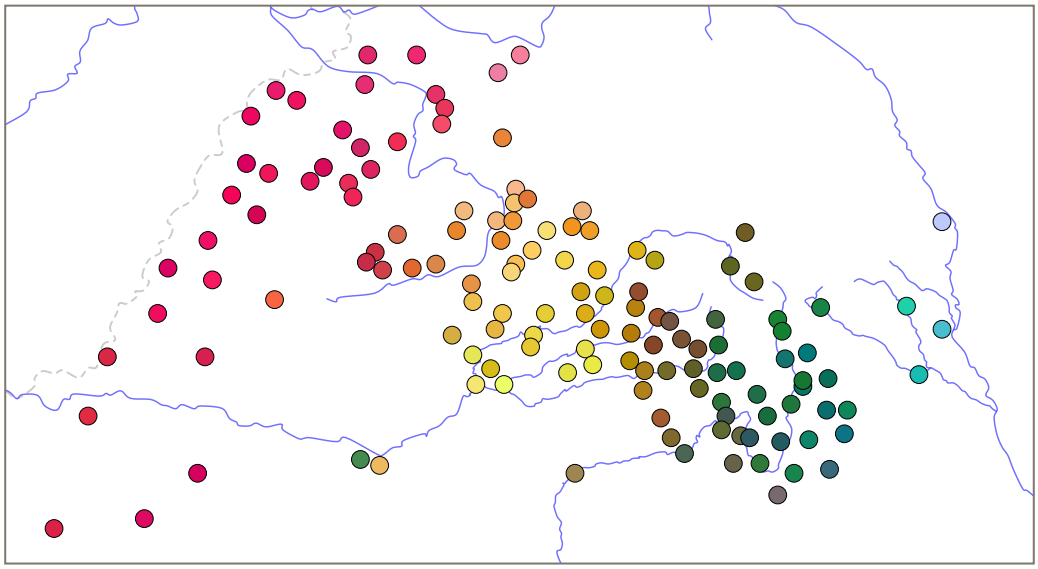
5.1. térkép: Az eredeti, finoman mellékjelezett lejegyzés mátrixának Ward-féle klaszteranalízise, 6 csoport térképezése esetén



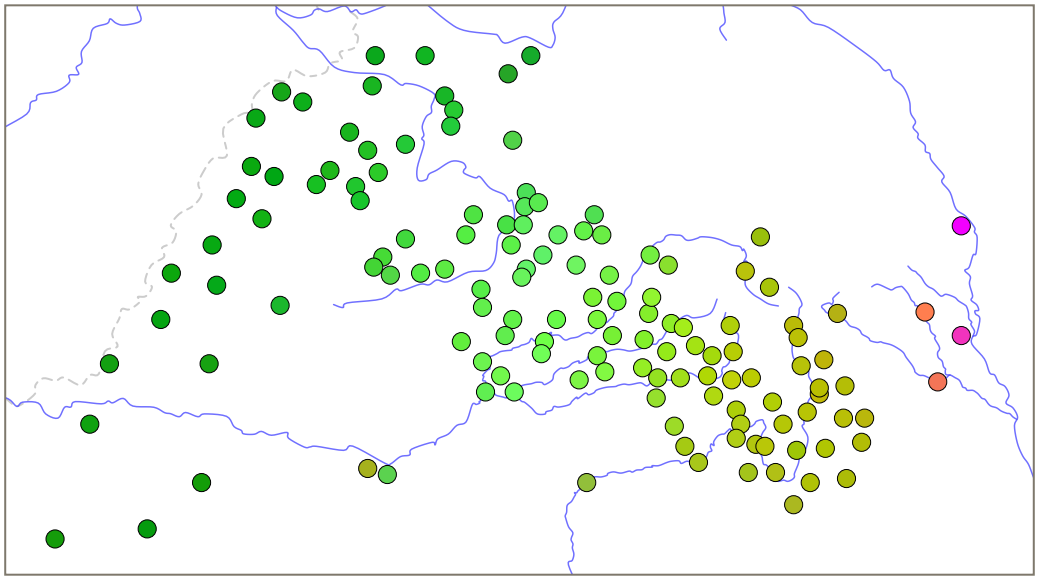
5.2. térkép: A lexikai mátrix Ward-féle klaszteranalízise, 6 csoport térképezése esetén

Öt csoport elkülönítése esetén (5.6. ábra) már különböző térképek jönnek létre, a fonetikailag érzékeny mátrix esetében nagyjából a Székelyföld középső részén alakul ki törésvonal, a lexikai mátrix esetében viszont a négy moldvai kutatópont válik ketté, szintén kelet-nyugati tagolódást mutatva. Hat csoport elkülönítése esetén (5.6. ábra) még különbözőbbek lesznek a klasztertékrépek: a fonetikai mátrix esetében a mezőségi kutatópontok válnak ketté, jól kivehető területi egységeket nem hozva létre, a lexikai mátrix esetében nagyjából Szamoshát és Szilágyság válik el a dél-nyugati kutatópon-

toktól. A fonetikai mátrixot a térképek alapján nem érdemes öt csoportnál többet mutató klasztertérképen ábrázolnunk.



5.3. térkép: A fonetikailag érzékeny mátrix többdimenziós skálázással készített térképe



5.4. térkép: A lexikai hangsúlyú mátrix többdimenziós skálázással készített térképe

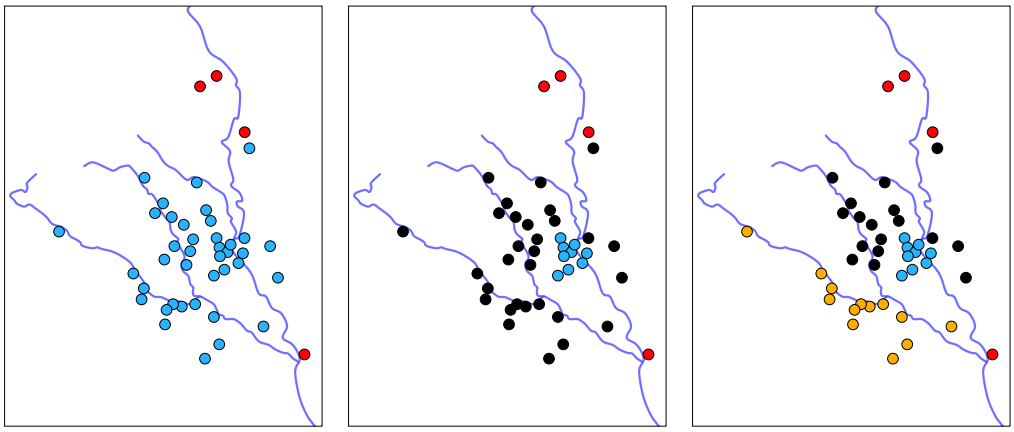
Összevetve a kalsztertérképeket és a többdimenziós skálázás eredményét megmutató 5.3. és 5.4. MDS-térképeket a következőket mondhatjuk. A fonetikailag érzékeny mátrix alapján jobban elkülönülnek egymástól a korábbi felosztások alapján kialakított nyelvjárásterületek kutatópontjai, nincs azonban éles határ, különösen a Székelyföld nyugati része és a mezőség között. A lexikai mátrix alapján még ennyire sem különülnek el egymástól az egyes területek, illetve a moldvai kutatópontok esetében láthatunk a Székelyföld keleti részétől számottevőbben elütő árnylatokat.

Összegezve a klaszter- és MDS-térképek tanulságait, a RMNyA. dialektometriai elemzése alapján legföljebb négy csoportot érdemes megkülönböztetnünk, amelyek a hagyományos felosztásnak nagyjából megfeleltethetők (Partium-Bánság, Mezőség, Székelyföld, Moldva). Az is látszik azonban, különösen a fonetikai mátrix MDS-térképén, hogy ezek a területek nem különülnek el egymástól élesen, és korántsem homogének. A Székelyföld nyugati fele például inkább átmenetet mutat az attól nyugatra fekvő Mezőség és a keleti székely területek között, a moldvai kutatópontok közül Szabófalva jobban elkülönül a másik három ponttól, amelyek azonban színárnyalatuk alapján (a fonetikailag érzékeny mátrix szerint) nyelvileg nincsenek annyira távol a keleti székely területektől.

5.4. A moldvai csángó nyelvjárás atlasza

A MCsNyA. az első, teljes egészében informatizált magyar nyelvjárási atlasz. A nyomtatásban megjelent első két kötet mellett az informatizált változat azoknak a térképlapoknak az adatait is tartalmazza, amelyek az adattár publikálása után kerültek elő (bővebben l. Bodó–Vargha 2008). A teljes adattárat táblázatos formában 2007-ben tettük közzé (Bodó–Vargha 2007). A MCsNyA. dialektometriai elemzéséről, különböző nyelvi hasonlósági mátrixainak összevetéséről e kötet korábbi fejezeteiben nem írok részletesen, erről lásd Bodó et al. 2012. és Bodó–Vargha 2016. Az alábbi térképek a MCsNyA. mátrixai klaszterelemzésének és többdimenziós skálázásának eredményét mutatják be röviden.

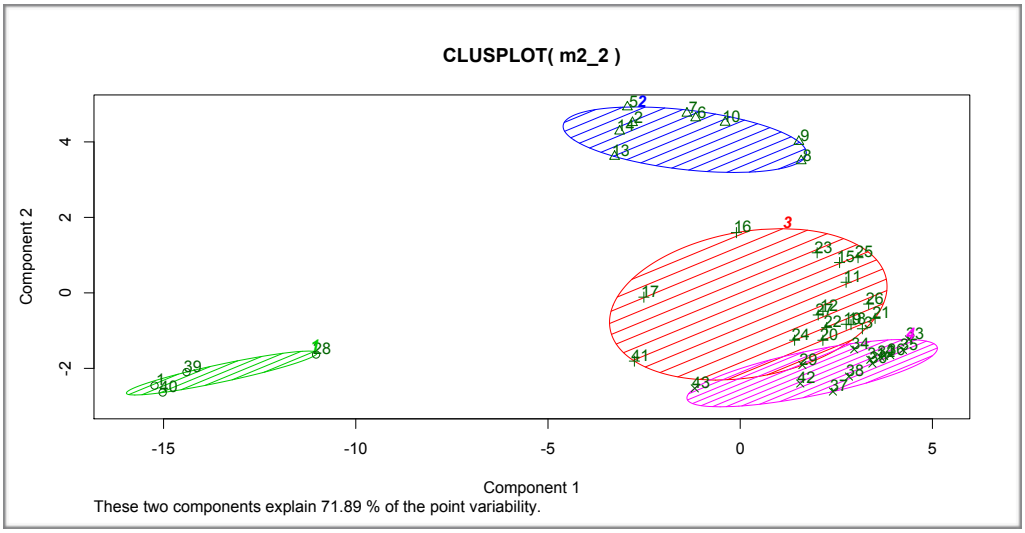
Az atlaszból a fonetikai részletekre is érzékeny mátrix alapján készítettem a kutatópontokat kettő, három és négy csoportra osztó klasztertérképeket. A kialakított csoportok közül az északi (a délre kirajzott Ploszkucényt is ide értve) különül el legmarkánsabban a többi kutatóponttól, amelyek között következő lépésben a Bákó környéki pontok különíthetők el, a második térképen fekete színnel jelölt csoport további tagolás esetén egy Tázló és egy Tatros menti területre válik ketté (l. még erről Bodó–Vargha 2016, vö. Péntek 2014).



5.7. ábra: A MCsNyA. eredeti, részletesen mellékjelezett lejegyzéséből készült mátrix Ward-féle klaszteranalízise, 2, 3 és 4 csoport térképezése esetén

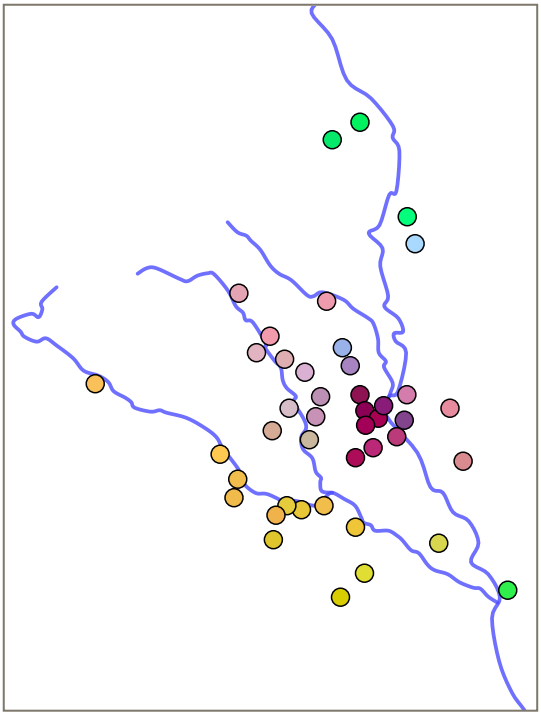
Azt, hogy mennyire relevánsak a klaszteranalízissel kialakított csoportok, további statisztikai módszerekkel vizsgálhatjuk. Főkomponens-elemzés segítségével ellenőrizhetjük, mennyire különülnek el egymástól a különböző csoportokba sorolt kutatópontok (a módszerről l. részletesen Ben-Hur–Guyon 2003). Az 5.8. ábra a

MCsNyA. fonetikailag érzékeny mátrixának főkomponens-elemzését mutatja a négy, Ward-féle klaszteranalízis alapján kialakított csoportot színekkel ábrázolva. Az 1. csoport (zöld színnel) a négy északi kutatópontot mutatja, amelyek egyértelműen elkülönülnek a többi kutatóponttól. A 2. (kék színű) csoport a Bákó környéki kutatópontokat tartalmazza, amelyek szintén világosan elválnak a Tázló és a Tatros menti kutatópontoktól. Ahogyan azt az 5.7. ábrán látható térképek alapján is valószínűsíthettük, a Tázló (3. csoport, pirossal) és Tatros menti (4. csoport, ciklámen színben) csoportok kevésbé élesen különülnek el egymástól, de az ábra alapján ennek a két csoportnak a megkülönböztetése is ésszerűnek mondható. Az is látszik azonban, hogy itt már nem minden kutatópont sorolható be egyértelműen valamelyik csoportba, Szászkút (29.) a 3. és a 4. csoport halmazának metszéspontjában helyezkedik el.



5.8. ábra: A MCsNyA. fonetikai mátrixának főkomponens-elemzése a klaszteranalízis alapján kialakított csoportosítás ábrázolásával

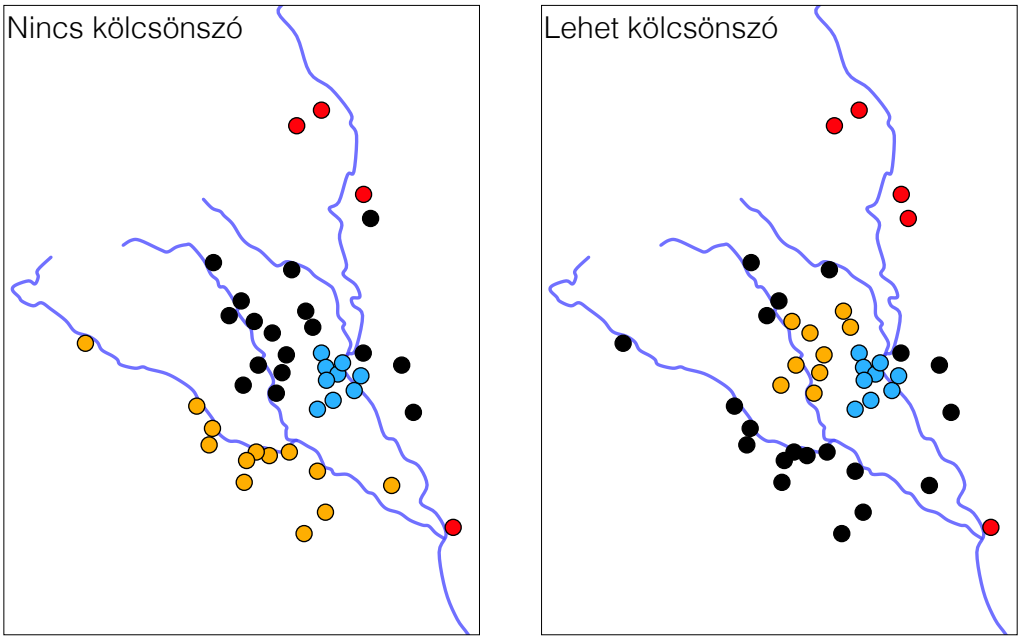
A MCsNyA. teljes anyagából, fonetikailag érzékeny mátrix alapján készült MDS-térkép a klaszteranalízis után tovább pontosítja a nyelvi hasonlósági mintázatokat, az átmeneteket és a négy kialakított csoporton belüli, kutatópontok közti hasonlóságot is jobban érzékeltetve. Ugyanakkor alá is támasztja a kutatópontok négyes felosztását. A színek alapján az északi a nyelvjárásilag leginkább homogén csoport, illetve a Tatros menti települések is szinte azonos színárnyalatot kapnak. A Bákó környéki kutatópontok esetében már nagyobb eltéréseket találunk, a leginkább heterogén a Tázló mente. Ha két kutatópontnak hasonló a színe, az nem feltétlenül azt jelenti, hogy egymásra hasonlítanak leginkább, hanem azt, hogy a többi kutatóponttal való hasonlósági értékeik hasonlóak. A térképen világoskék Kalugarény és Szakatura például 68%-ban hasonlít egymásra, azonban mindkét kutatópont relatíve (az atlasz kutatópontjai között hasonlósági sorrendet felállítva) nagyobb mértékű hasonlóságot mutat a Tázló keleti partján lévő kutatópontokkal, úgy mint Szerbek, Esztrugár, Gajdár, Esztufuj. A hasonlóság százalékban kifejezett mértéke minden esetben 70% fölötti, Szakatura esetében azonban, amely a kérdéses kutatópontokhoz földrajzilag közelebb van, ez az érték mindig 1-2%-kal nagyobb, mint Kalugarénynál.



5.5. térkép: A MCsNyA. eredeti, részletesen mellékjelezett lejegyzéséből készült mátrix többdimenziós skálázásának térképe

Dialektometriai elemzést végezhetünk úgy is, hogy valamilyen szempontból csoportosítjuk a térképlapokat, és az egyes csoportokról külön készítünk hasonlósági mátrixot. Korábbi kutatásunkban (Bodó–Vargha 2016) azt vizsgáltuk, milyen hatással lehetnek a moldvai nyelvjárásokban jellemzően újabb elemként meglévő román kölcsönszók a MCsNyA. kutatópontjai közötti nyelvi hasonlósági viszonyokra. Annak érdekében, hogy a kölcsönszói hatást megragadhassuk, két csoportba soroltuk a térképlapokat aszerint, tartalmaznak-e román kölcsönszót, avagy sem. A térképek besorolását Bodó 2007b-es elemzése alapján készítettük. Az így kialakított alkorpuszokból fonetikailag érzékeny, illetve lexikai szintű különbségeket tükröző nyelvi hasonlósági mátrixokat készítettünk.

Az eredmények alapján a román kölcsönszók egyik nyelvjárásból a másikba is átkerülhetnek, közvetlen román kontaktus nélkül is, a lexikon szintjén megváltoztatva a kutatópontok közti hasonlósági viszonyokat, a földrajzi közelség függvényében (bővebben l. Bodó–Vargha 2016). Így a kölcsönszókat nem tartalmazó és a kölcsönszókat tartalmazó térképek mátrixa alapján eltérő nyelvi hasonlósági mintázatok rajzolódnak elénk (l. 5.9. ábra). A különbség leginkább az északi Balanyászával szomszédos Kalugarényt, illetve a Bákó környéki területtől nyugatra lévő Tázló menti településeket érinti, vagyis olyan, az 5.7. ábrán a Tázló menti csoporthoz tartozó településeket, ahol a csoporton belül nagyobb a kölcsönszók előfordulási gyakorisága az adatokban (vö. Bodó 2007b). Így a kölcsönszókat is tartalmazó térképlapok lexikai mátrixának klaszteranalízise alapján másképpen tagolódnak a kutatópontok.



5.9. ábra: A MCsNyA. kutatópontjainak 4 csoportba sorolása, kölcsönszókat nem tartalmazó, illetve kölcsönszókat tartalmazó térképlapok lexikai hangsúlyú mátrixai alapján

A kölcsönszók eszerint a korábbi, a moldvai magyar falvak eredeti nyelvjárásához köthető hasonlósági mintázatokat részlegesen elfedhetik. E megállapításnak különösen akkor van jelentősége, ha történeti kapcsolatok feltárása céljából integrált dialektometriai elemzést végzünk a MCsNyA. és egy másik nyelvatlasz, például a RMNyA. anyagán. Moldva más területekkel való nyelvjárási kapcsolatainak vizsgálatához célszerű olyan térképlapokkal dolgoznunk, amelyek nem tartalmaznak kölcsönszókat.

5.5. A romániai magyar nyelvjárások atlasza és A moldvai csángó nyelvjárás atlasza integrált elemzése

A RMNyA. és a MCsNyA. integrált dialektometriai elemzését a MCsNyA.-ban a román kölcsönszók nyelvi hasonlósági mintázatokra gyakorolt hatását vizsgáló elemzés tanulságait felhasználva készítettem el (l. Bodó–Vargha 2016, illetve az 5.4. alfejezetet). Ennek értelmében kizárólag olyan térképeket használtam föl, amelyek az MCsNyA.-ban nem tartalmaznak kölcsönszót. A megfelelő térképlapokat egy algoritmus segítségével válogattam ki a két adattárból, ahol a kiválasztás feltétele az volt, hogy a térképek címszava azonos legyen, és a MCsNyA. térképlapján kölcsönszónak minősített adatok ne szerepeljenek. Így összesen 266 RMNyA. és értelemszerűen ugyanennyi MCsNyA. térképlapon tudtam elvégezni a nyelvi hasonlósági mátrix kialakításához szükséges adatösszevetéseket.

A térképlapokból két nyelvi hasonlósági mátrixot hoztam létre. Az egyik az eredeti lejegyzés alapján készült, így érzékeny az apróbb fonetikai különbségekre is, a másik a lejegyzés olyan, radikálisan egyszerűsített változata alapján készült, amely többnyire már csak a lexikai szintű különbségeket teszi megragadhatóvá. Így a két mátrix alapján készített térképeket a több térképet összesítő ábrákon FON (fonetikai különbségekre is érzékeny) és LEX (lexikai szintű különbségeken alapuló) jelöléssel láttam el.

Az 5.10. ábra a RMNyA. és a MCsNyA. integrált dialektometriai elemzéséről készített Ward-féle klaszteranalízis eredményét mutatja kettő és három csoport térképezése esetén. A FON mátrix esetében két csoport megjelenítése esetén a „határ" a Székelyföld és Mezőség között húzódik, vagyis a moldvai régió és a Székelyföld egy csoportba kerül. A LEX mátrix esetében a MCsNyA. kutatópontjai különülnek el a RMNyA. kutatópontjaitól. Három csoport megkülönböztetése esetén hasonló térképeket kapunk, a FON térképen a keleti rész ketté válik, a LEX térképen pedig a Mezőség és a Székelyföld (ide értve az annak nyugati szegélyén elhelyezkedő kutatópontokat) között válnak ketté a színek.

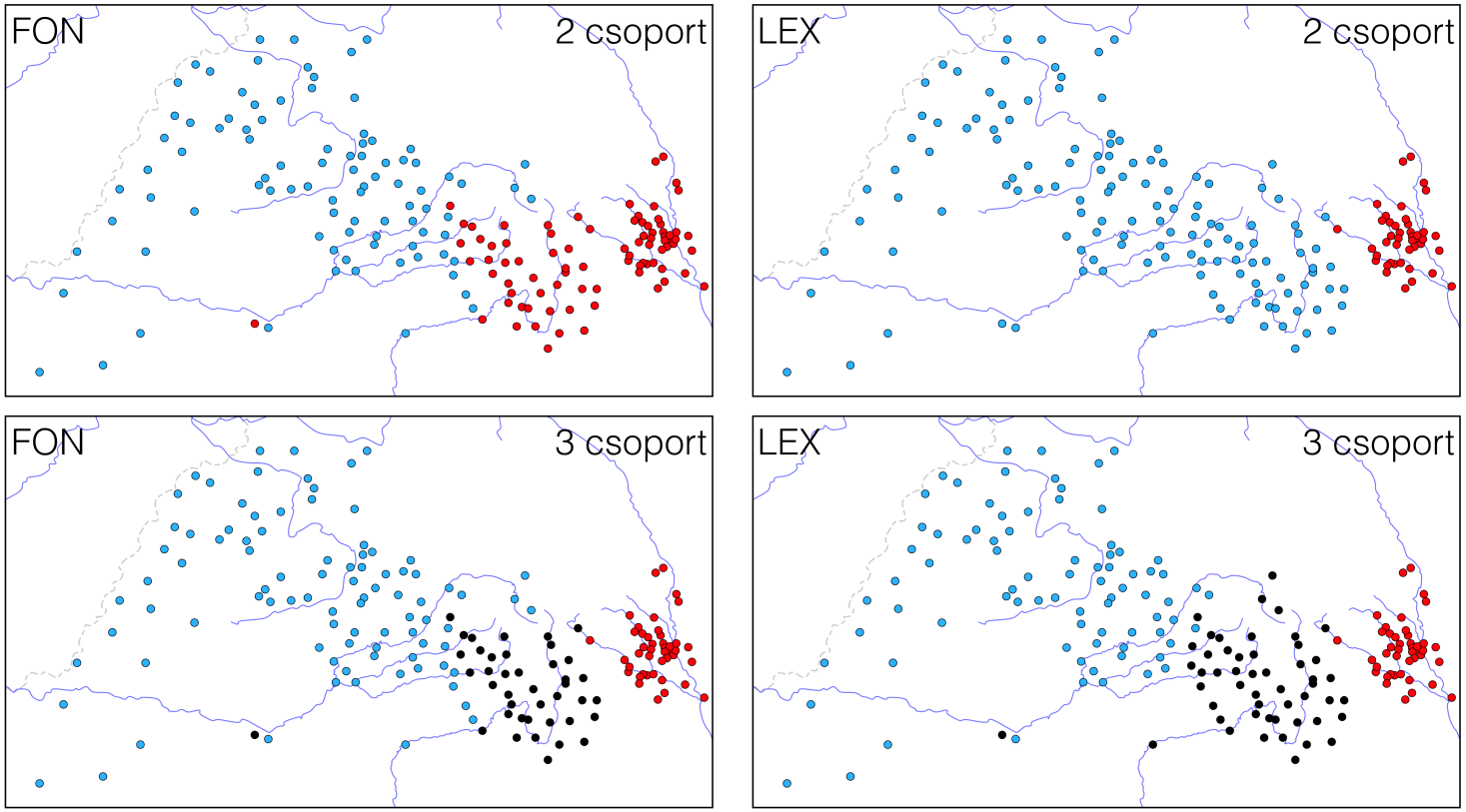
Négy és öt csoport megkülönböztetése is hasonló eredményekkel jár a két mátrix esetében. Először nagyjából a partiumi és a mezőségi kutatópontok kerülnek külön csoportba, itt annyi a különbség, hogy a Mezőség fölött, északon található három kutatópont a FON mátrix esetében a nyugati, a LEX mátrix esetében a keleti csoportba sorolódik be. Öt csoport megkülönböztetése esetén a MCsNyA. kutatópontjai válnak ketté, külön csoportba kerülnek az északi kutatópontok.

Hat csoport megjelenítésekor a közbülső, leginkább a Mezőségnek megfeleltethető területen a FON mátrix esetében kétfelé válnak a kutatópontok, de a két csoport nem ad ki egyértelmű területi egységeket. A LEX mátrix hat csoportot megmutató klasztertérképén a MCsNyA. kutatópontjai bomlanak tovább, immár három csoportra.

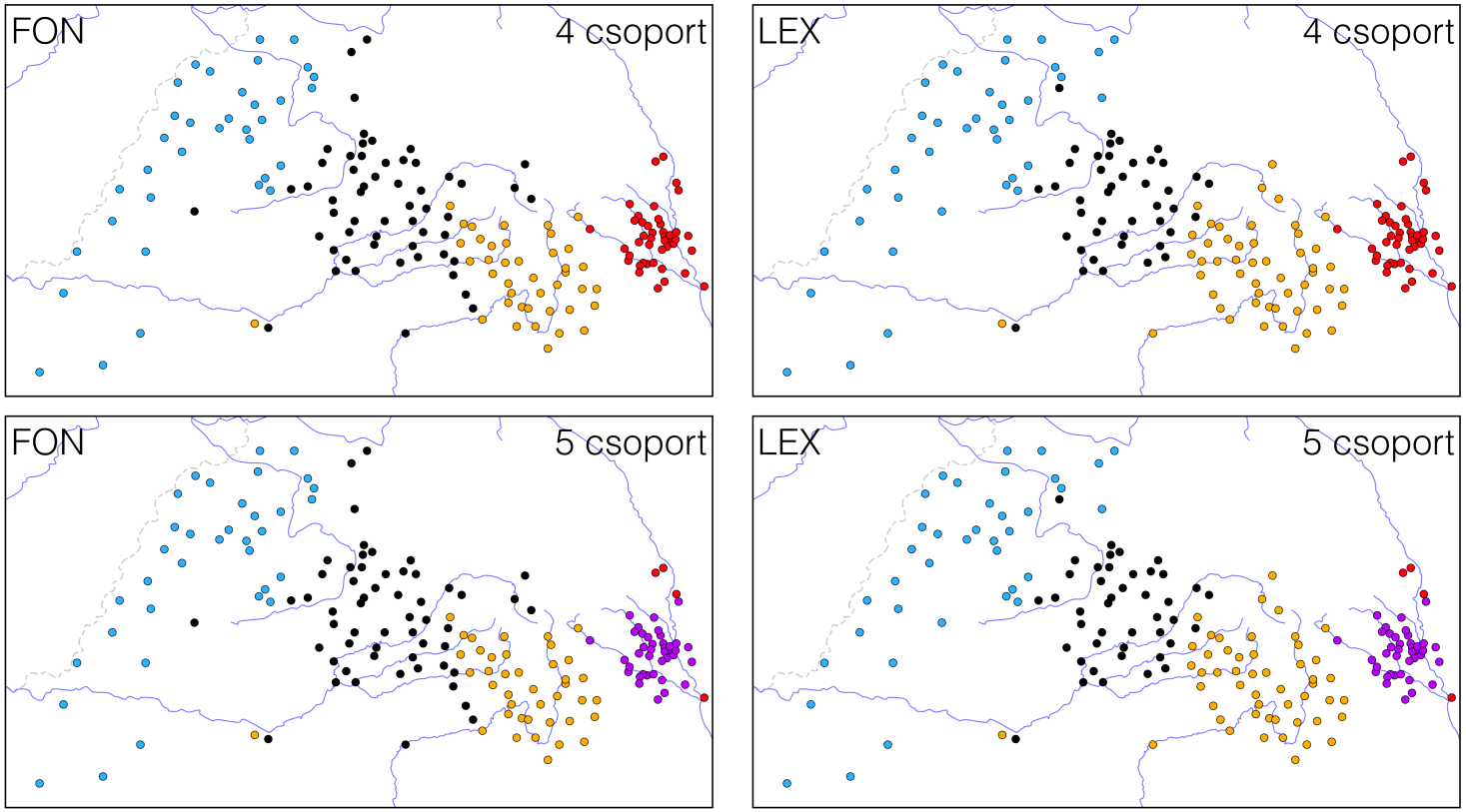
A klasztertérképek alapján mindkét mátrixot figyelembe véve négy, esetleg öt nagyobb területi egységet érdemes megkülönböztetnünk: Partium, Mezőség, Székelyföld, Moldva. Míg azonban a LEX mátrix esetében Moldva élesebben elkülönül a többi kutatóponttól, addig annak ellenére, hogy az adatok nem ugyanabból az adattárból származnak, a FON mátrix alapján Székelyföld és Moldva egy csoportba kerül a kutatópontok kettős felosztásakor.

Érdemes összevetnünk a klasztertérképeken kirajzolódó csoportokat a két mátrixból készített MDS-térképekkel. Mindkét térképet, hiszen nem kategorikus felosztásról van szó, az átmenetek jellemzik, a FON mátrix esetében inkább látszik valamiféle, a klasztertérképekre emlékeztető területi felosztás, illetve a MCsNyA. területén az északi kutatópontok zöldes árnyalata egyszerre mutatja a négy kutatópont elkülönülését a többi moldvai kutatóponttól és érzékelteti e települések és az észak-mezőségi kutatópontok nyelvi hasonlósági mintázatainak párhuzamait. A LEX mátrix MDS-térképén inkább a földrajzi távolság szerint alakulnak a színek, nyugat-keleti irányú kontinuumot látunk, Moldvában az északi kutatópontok színe a Bákó környéki kutatópontokéra hasonlít leginkább, az észak-mezőségiekre nem.

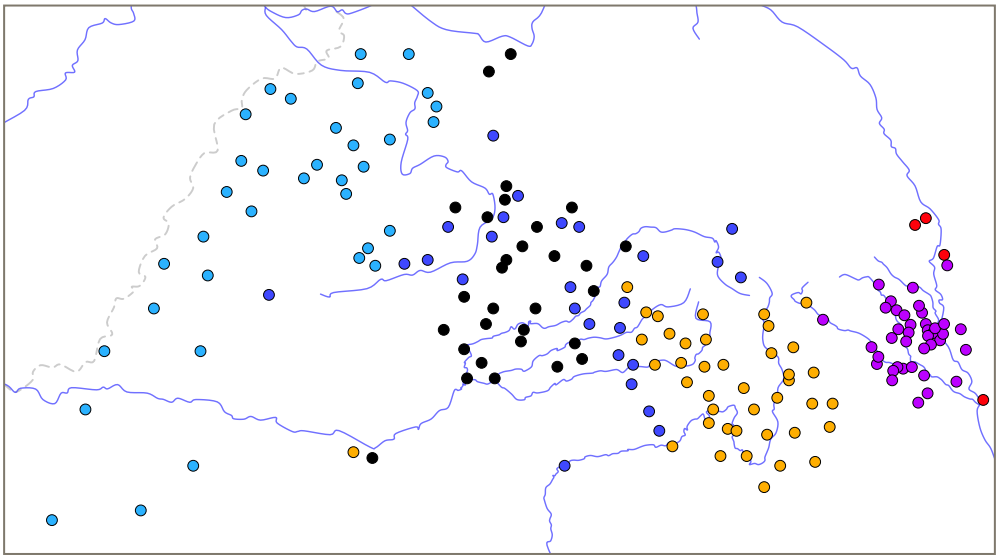
Bár mindkét adattár, a RMNyA. és a MCsNyA. is az erdélyi lejegyzési hagyományokat követve készült, nem zárhatjuk ki, hogy itt is felléphet adattár-hatás, mint ahogyan azt a MNyA. és a RMNyA. integrált elemzése kapcsán láthattuk (l. 4.2. alfejezet). Elképzelhető, hogy a lejegyzés összehangolásával nagyobb lenne a hasonlóság a keleti székely és a Tatros, illetve a Tázló menti kutatópontok között. Moldva és a Székelyföld nyelvi hasonlósági viszonyainak vizsgálatához a Székely nyelvjárási atlasz és a MCsNyA. integrálásával juthatnánk igazán közelebb, a Székely nyelvjárási atlasz azonban még nagyrészt informatizálásra vár (a háromszéki rész térképeinek informatizálásáról l. Cs. Nagy 2011, illetve megkezdődött a cédulaanyag feldolgozása is, l. Both, megj. előtt).



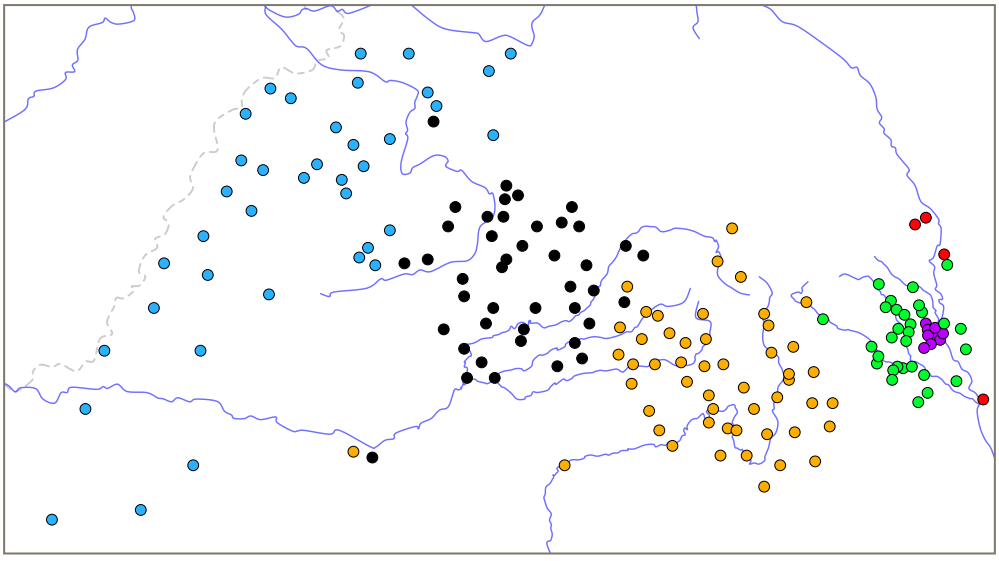
5.10. ábra: A RMNyA. és az MCsNyA. integrált dialektometriai elemzésének Ward-féle klaszteranalízises térképei különböző mátrixok alapján, 2 és 3 csoport térképezése esetén



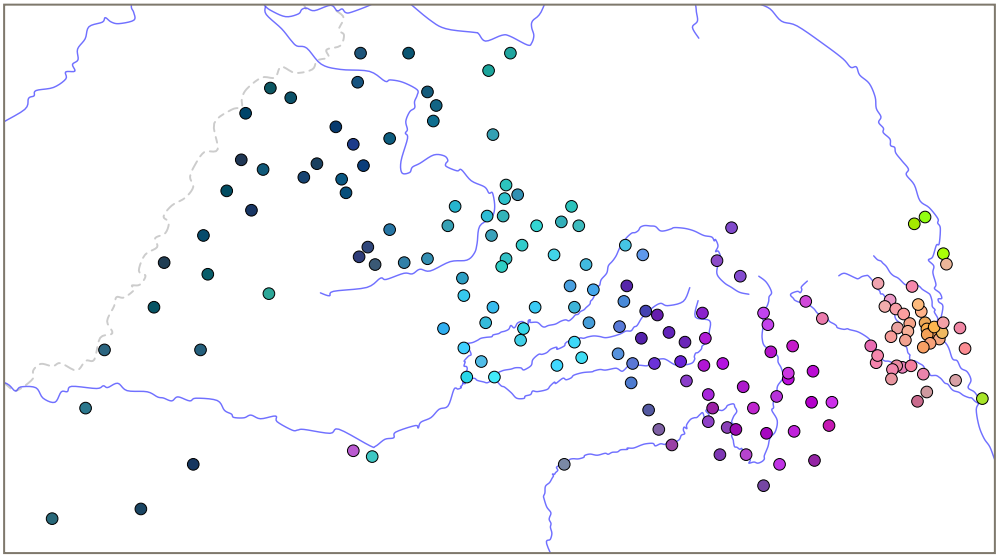
5.11. ábra: A RMNyA. és az MCsNyA. integrált dialektometriai elemzésének Ward-féle klaszteranalízises térképei különböző mátrixok alapján, 2 és 3 csoport térképezése esetén



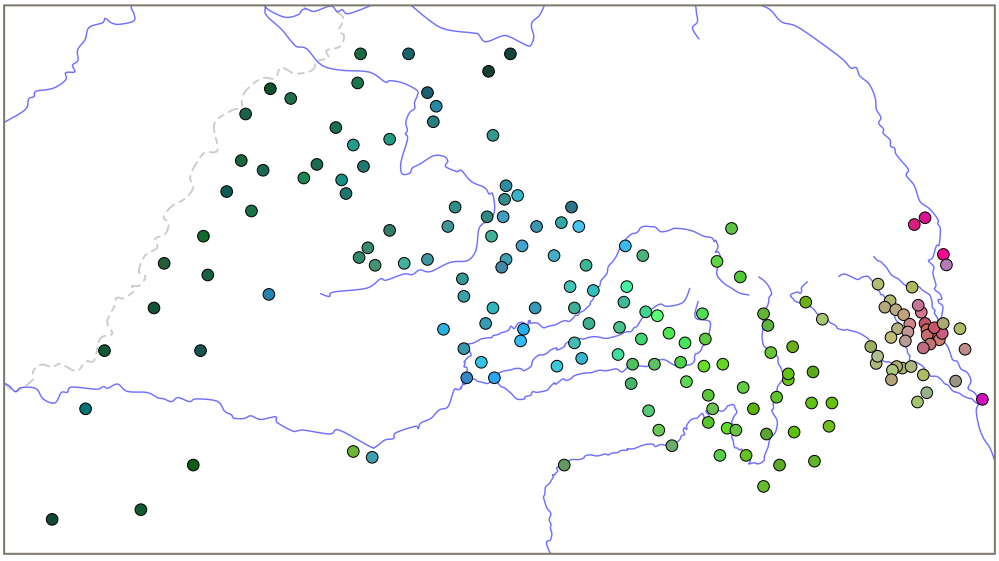
5.6. térkép: A RMNyA. és a MCsNyA. kutatópontjainak felosztása 6 csoportra a fonetikailag érzékeny mátrix Ward-féle klaszteranalízise alapján



5.7. térkép: A RMNyA. és a MCsNyA. kutatópontjainak felosztása 6 csoportra a lexikai mátrix Ward-féle klaszteranalízise alapján



5.8. térkép: A RMNyA. és a MCsNyA. fonetikailag érzékeny mátrixából készített többdimenziós skálázás eredményének térképe



5.9. térkép: A RMNyA. és a MCsNyA. lexikai hangsúlyú mátrixából készített többdimenziós skálázás eredményének térképe

5.6. A magyar nyelvjárások felosztása – A magyar nyelvjárások atlasza és A romániai magyar nyelvjárások atlasza integrált elemzése

Nincs a teljes magyar nyelvterületet lefedő nyelvjárási adattárunk; ha az Őrvidéktől Moldváig szeretnénk dialektometriai alapon felosztani a nyelvjárásokat, a MNyA. mellett szükségünk van a RMNyA.-ra is. A 4. fejezetben többféle elemzés alapján is vizsgáltuk a két atlasz kutatópontjainak nyelvi hasonlósági mintázatait, illetve össze is vetettük egymással a különböző mátrixokat. Összesen négy mátrixszal dolgoztunk: az első az eredeti, finoman mellékjelezett lejegyzés alapján készült, a második 127 térképlap kutatói csoportosításából, a harmadik a különböző adattárakból származó adatok integrálása szempontjából optimális fonetikai pontosságú lejegyzés szerint, a negyedik pedig a lexikai szintű különbségek megragadására alkalmas, egyszerűsített lejegyzés alapján. Az itt bemutatott térképek is e négy mátrix alapján készültek.

Kilencszer négy, összesen tehát harminchat térkép alapján mutatom be, milyen eredményeket hoz a Ward-féle klaszteranalízis a különböző mátrixok használatakor a csoportok számának függvényében. A klaszteranalízis által kialakított csoportosítást kettőtől tízig (ez kilenc eset) jelenítettem meg a térképeken, azért választottam a tízes megjelenítési mélységet maximumnak, mert a magyar nyelvterületet régiókra bontó legújabb térkép is tízes felosztást mutat (Juhász 2001b). További térképeken látható, milyen eredményeket hoz a többdimenziós skálázás a négy különböző mátrix esetében.

A színek alapján a térképek jól összevethetők, érdemes azonban néhány lényeges különbségre fölhívnunk a figyelmet. Már a kutatópontok két csoportba sorolásakor is (5.12. ábra) jelentős különbséget látunk a különböző mátrixok között. A két, fonetikai különbségekre érzékeny elemzés (l. a két fölső térképet) nagy vonalakban a Tisza mentén, míg a csoportosítás alapján készült és a lexikai hangsúlyú mátrix szerint a Duna mentén válnak inkább ketté a kutatópontok. Ha négy csoportot jelenítünk meg (5.14. ábra), még markánsabb eltéréseket figyelhetünk meg: a fonetikai mátrixok klasztertérképein jól elkülönül a Nyugat-Dunántúl és a palóc nyelvjárási terület, míg a másik két térkép inkább a földrajzi távolság alapján, a Dunántúlt egységesnek mutatva osztja meg a kutatópontokat. A korábbi fejezetekben láthattuk, hogy egyes – például megyehatáron, illetve a Palócföld nyugati szélén lévő vagy sziget helyzetű – kutatópontok nyelvi hasonlósági mintázatai jelentősen különbözhetnek az elemzés hátterében lévő lejegyzés fonetikai pontossága szerint. A kutatópontok többségénél ugyan nem tapasztalunk jelentős eltérést a földrajzi hasonlósági viszonyokban, az automatikus csoportosítás eredménye azonban még ezen kutatópontok esetében is igen elütő lehet a mátrix függvényében.

Megállapítható tehát, hogy kettő, három és négy csoport elkülönítése esetén a csoportosítás alapján készített mátrix nem a fonetikai, hanem a lexikai jellegű elemzéssel mutat inkább hasonlóságot, holott a kialakított munkatérképek túlnyomó többsége fonetikai jellegzetességek földrajzi megoszlását mutatja meg (l. részletesen a 4. fejezetben). Mi az, ami a csoportosításon alapuló mátrixból mégis hiányzik, és jellegében a lexikai hangsúlyú mátrixhoz teszi hasonlatossá? Már a 4.3. alfejezetben láthattuk, hogy a csoportosítás alapján készült mátrix kevésbé alkalmas a Palócföld határának megmutatására, illetve székely települések dunántúli nyelvi kapcsolatainak érzékeltetésére, mint a finom fonetikai. A legkézenfekvőbb magyarázat, hogy bár minden, a nyelvjárások leírásában, klasszifikációjában szerepet játszó hangtani jelenség helyet kapott a munkatérképek kialakításakor, az egyes jelenségek előfordulási gyakoriságát nem vettük, nem vehettük figyelembe. A nagy hatókörű, gyakori fonetikai jelenségeknek eszerint alapvető fontossága van mind a nyelvjárási törésvonalak meghatározásában, mind a nyelvjárások közti történeti kapcsolatok feltárásában.

Az iménti megállapítás a korábbiakhoz képest más megvilágításba helyezi az egyes nyelvi jelenségek szerepét a történeti összefüggések feltárásában. A nyelvföldrajzi jelenségek történeti célú vizsgálatának alapja elsősorban Benkő Loránd 1966-ban elhangzott akadémiai székfoglalója (vö. Juhász 2007). Annak tárgyalásakor, milyen jelenségeket érdemes figyelembe vennünk a nyelvjárásközi összevető vizsgálatokban, Benkő a következőképpen fogalmaz: „A hangtan keretében például viszonylag kevésbé jöhetnek számba az olyan, a hangrendszerben széles elterjedtségű, nagy hatóerejű jelenségek, mint például az *í*-zés, *á*-zás, diftongizáció, palatalizáció stb., noha más természetű jelenségekkel együtt járva természetesen ezek vallomásának is lehet bizonyos súlya. Sokkal nagyobb figyelem illeti ellenben tárgyunk szempontjából a hangrendszerben elszigeteltebb jellegű, kis hatóerejű jelenségeket. Például *körte* szavunk *körfe* változata (megvan az Ormánságban és Dél-Gömörben), vagy a *szégyell* ige *szégell* változata (megvan a Délnyugat-Dunántúlon, Pozsony-Nyitra vidékén, valamint néhány dél-erdélyi nyelvjárásszigeten) nagyobb valószínűséggel zárható ki az első [megőrzött régiség, archaizmus] és a második [egymástól függetlenül keletkezett neologizmus] magyarázati lehetőségből, mint a harmadikból [népmozgalmi okokkal magyarázható egyezés].” (Benkő 1967: 39)

A csoportosítás alapján létrehozott mátrix tanulságaiból leszűrhető azonban, hogy a nagy hatókörű jelenségek figyelembevétele valójában csak úgy valósulhat meg, ha nagy számú adattal dolgozunk, teljes körű, minden nyelvi jelenségre kiterjedő, a gyakorisági arányoknak megfelelő elemzést végzünk. Ezt csak a dialektometria módszertanával tudjuk megvalósítani, Benkő viszont még csak a nagyatlasz kérdőfüzeteinek áttekintésével dolgozhatott, kézenfekvőbb volt tehát számára, hogy néhány jól kiválasztott címszóra alapozza gondolatmenetét.

Megjegyzendő, hogy a Benkő által a hangrendszerben elszigeteltebb jellegű, kis hatóerejű jelenségeknek nevezett példák valójában lexikai kötöttségűek, mert bár egy-egy hangot érintenek az adatok közti különbségek, a megfelelések a hangrendszer szint-jén nem értelmezhetők, legföljebb néhány lexémát érintenek, csakúgy, mint például az amerikai angolban a szóhangsúly szintjén egyes lexémákban jelentkező különbségek, amelyeknek az elterjedtsége jellegzetesen különbözik a magánhangzórendszer sajátságainak térbeli mintázataitól (vö. Dinkin–Evanini 2010).

Négynél több csoport elkülönítésekor (5.15–5.20. ábra) a csoportosítás alapján készült mátrix klasztertérképe már kevésbé feleltethető meg a lexikai mátrixénak, inkább a főbb fonetikai jellegzetességeket figyelembe vevő, egyszerűsített lejegyzésből készült térképre hasonlít, azzal az igen lényeges eltéréssel, hogy a palóc nyelvjárások nem alkotnak összefüggő egységet.

A lexikai mátrix klasztertérképe leginkább a földrajzi távolságok mentén alakul, vagyis a kialakuló „régiók” jellemzően hosszanti csíkokat alkotnak, nyugattól kelet felé haladva. A Dunántúl itt a legegységesebb, nyolc csoport megkülönböztetésekor válik csak ketté.

A magyar nyelvjárásterület régiókra bontásának (Juhász 2001b) leginkább az egyszerűsített lejegyzés klasztertérképe feleltethető meg (l. 5.20. ábra), de itt is vannak lényeges különbségek, főként délen, illetve a Tiszától keletre. A moldvai kutatópontok nem alkotnak önálló csoportot egyik térképen sem, az egyszerűsített lejegyzés és a lexikai mátrix klasztertérképein Szabófalva és Bogdánfalva a mezőségi kutatópontok csoportjába (5.16–5.18 ábra), illetve külön csoportba kerül (5.15. ábrától), Pusztina és Diószeg azonban minden térképen a székelyföldi kutatópontokhoz tartozik.

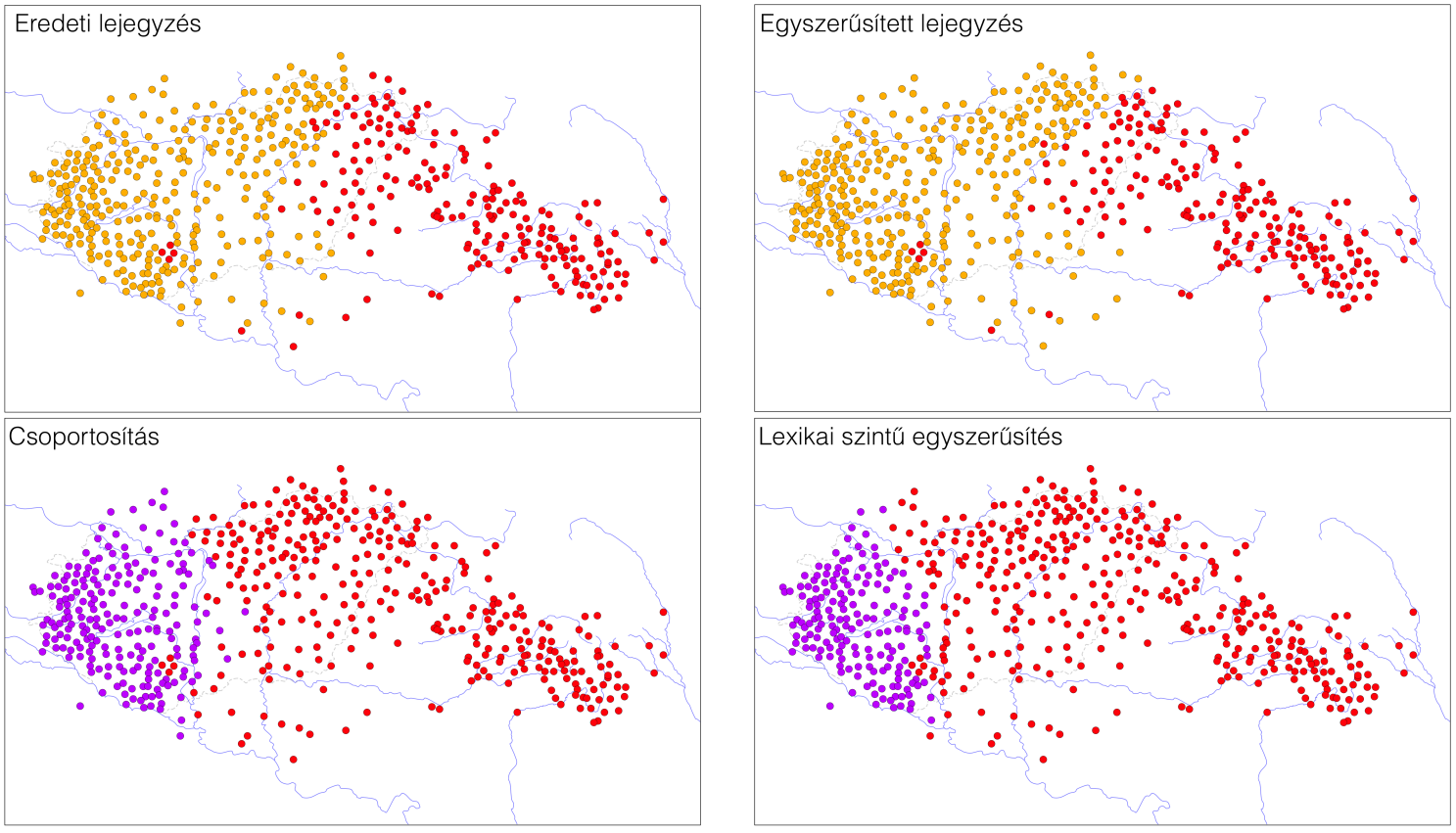
Az itt bemutatotthoz igen hasonló módszertanú elemzéssel rendelkezünk a Magyar néprajzi atlaszról is (Borsos 2011). Részletes összevetésre itt nem vállalkozhatom

(ehhez alapos, a kutatópontokat egymásnak megfeleltető, a mátrixokat statisztikai módszerekkel összevető elemzésre volna szükség), a klaszteranalízisek eredményét szemléltető néprajzi térképek alapján annyit mégis érdemes megjegyezni, hogy közülük a néprajzi atlasz nyelvi anyagának klaszterezése mutat leginkább hasonlóságot a nyelvatlaszok elemzésével, mégpedig leginkább a csoportosítás alapján készült mátrix klasztertérképével. A kilenc csoportot mutató térképek szinte teljesen azonosak, azzal a különbséggel, hogy a néprajzi térképen a tiszántúli kutatópontok oszlanak több csoportra, a csoportosítás alapján készült nyelvi térképen az erdélyi kutatópontok mutatnak megoszlást (vö. Borsos 2011: II./61). A térképek közötti feltűnő hasonlóság leginkább az elemzési módszerek közti párhuzammal magyarázható, illetve azzal, hogy a vizsgált jelenségek jellegéből adódóan a néprajzi térképen sem érvényesül a nagy hatókörű hangtani jelenségek hatása.

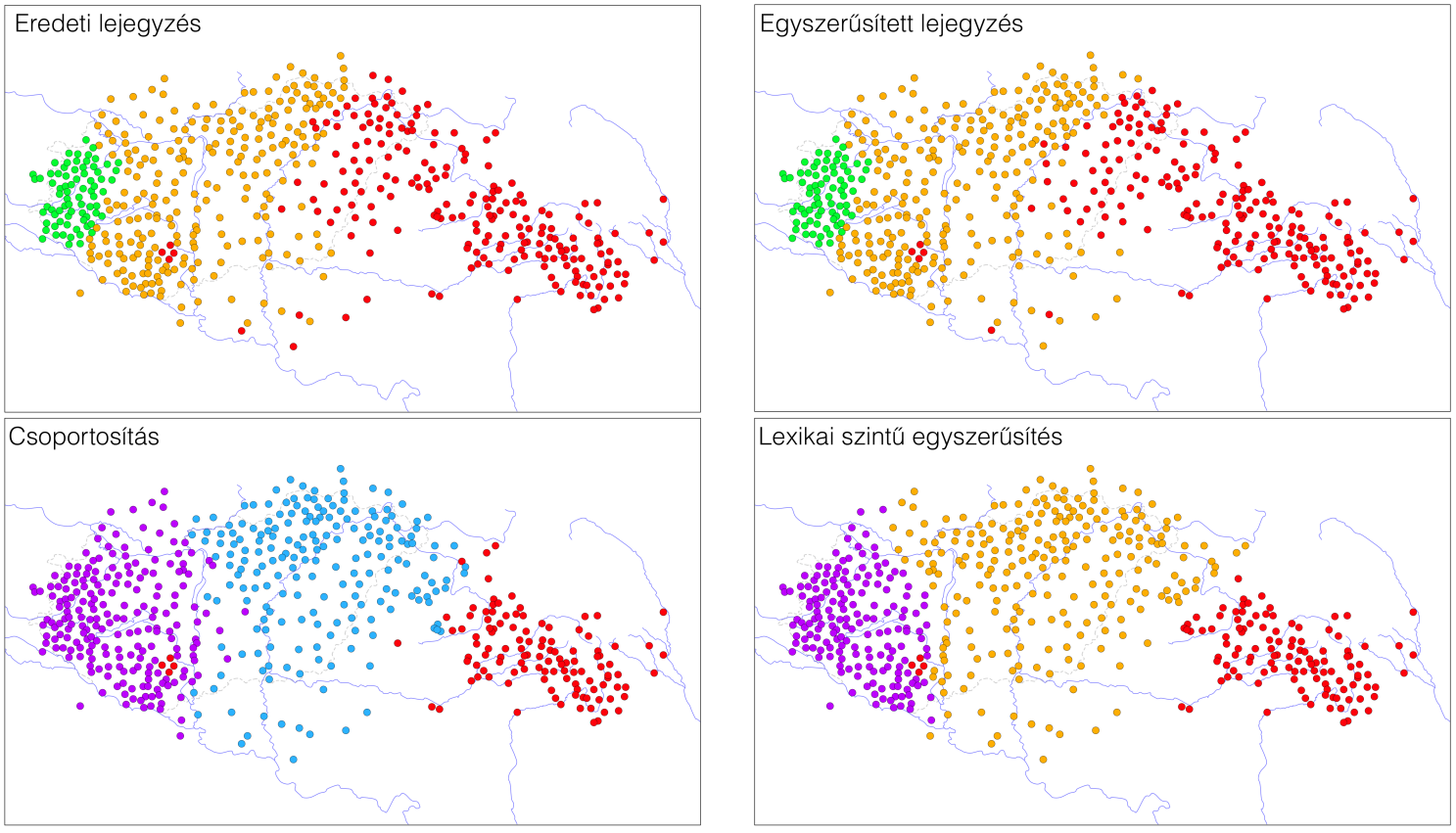
Érdemes még odafigyelnünk arra is, hogyan jelentkezik a térképeken az adattár-hatás, vagyis mennyire látszik meghatározónak a csoportok kialakításában az, hogy melyik adattárhoz tartoznak a kutatópontok. Esetünkben az adattárak közti határ egybeesik az országhatárral, így azt érdemes figyelni, mely térképeken kapnak más színt a kutatópontok éppen a határ mentén. A lexikai mátrix esetében nyolc, az eredeti lejegyzés alapján készített mátrix esetében kilenc csoport megkülönböztetésekor válnak ketté a kutatópontok a határ mentén. A finoman mellékjelezett lejegyzés esetében ezt a mellékjelezésben, illetve egyes hangok (pl. diftongusok) lejegyzési gyakorlatában lévő különbségek indokolják. Az adattár-hatás jelenlétére a lexikon szintjén más magyarázatot kell keresnünk. Elképzelhető, hogy az országhatár néhány évtized alatt is hatással lehet a hangtaniaknál jóval gyorsabb lexikai változásokra, de az is elképzelhető, hogy a terepmunka módszertanából, néhány eltérően megfogalmazott kérdésből adódik a különbség. A kérdés megválaszolásához mindenképp további, kvalitatív vizsgálatok szükségesek, amelyek figyelembe veszik a két adattár közti módszertani, adatközlési különbségeket.

Az egyszerűsített lejegyzés alapján készített, illetve a csoportosításon alapuló mátrix esetében kevésbé érvényesül az adattár-hatás, a csoportok nem a határ mentén válnak ketté, tíz csoport megkülönböztetésekor sem.

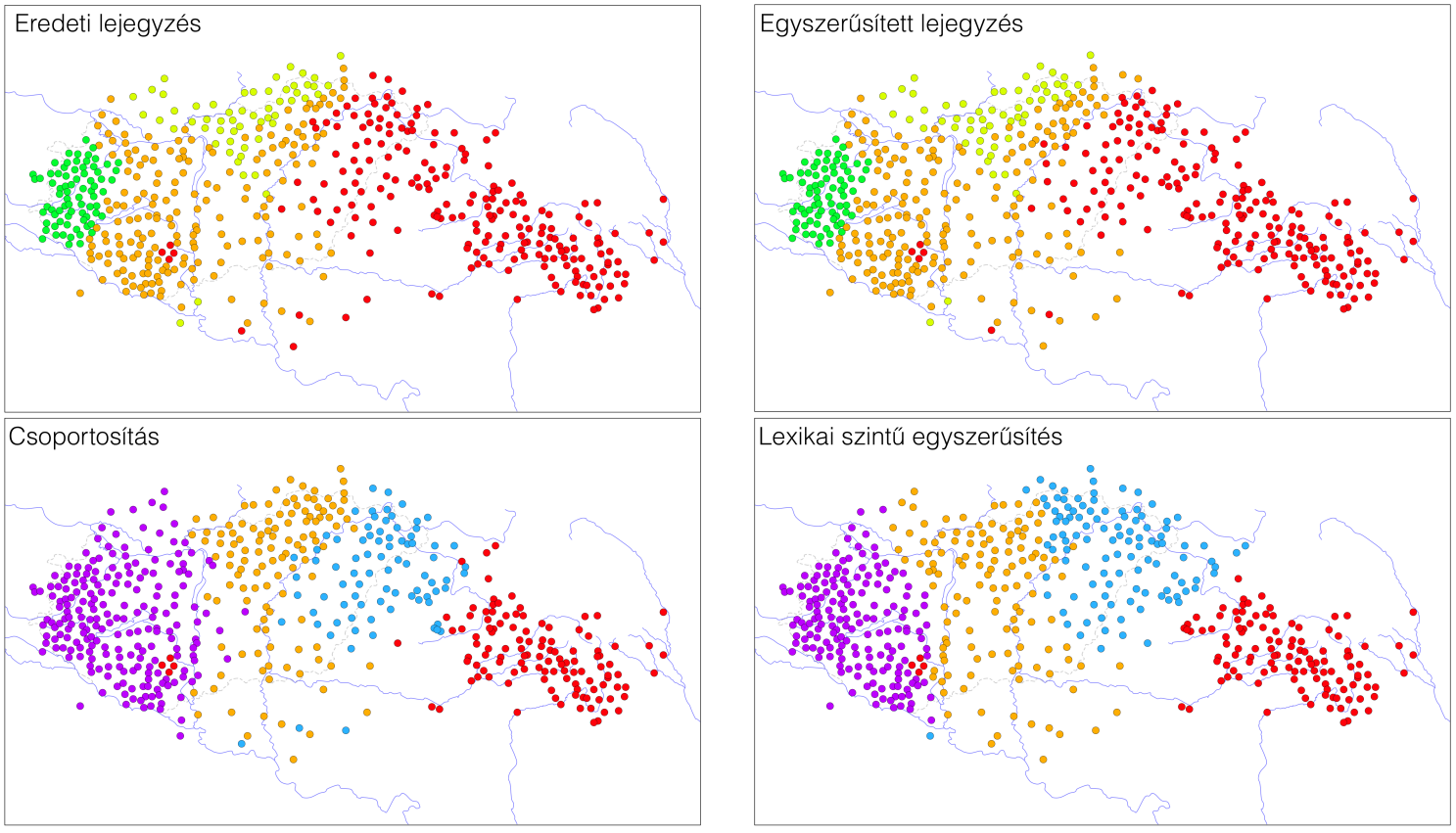
A nyelvjárások közti hasonlósági viszonyok alakulását kontinuumszerűen láthatjuk a térbeli összefüggéseket többdimenziós skálázás eredményeként bemutató 5.10, 5.11, 5.12 és 5.13 térképen. Az eredeti és a valamelyest egyszerűsített lejegyzés alapján készült térképeken, jóllehet a színek árnyalatonként változnak, jól látszik a Nyugat-Dunántúl és a palóc nyelvjárások elkülönülése is. A csoportosítás alapján készült térképen kevésbé élesen látszik kirajzolódni a Nyugat-Dunántúl, látszik azonban annak észak-déli tagolódása, a palóc nyelvjárások nem válnak ki a környezetükből, viszont ezen a térképen van a legélesebb különbség a történelmi Erdély határvonalában a kutatópontok színe között. A lexikai mátrix teljes egészében kelet-nyugati irányú átmenetet mutat.



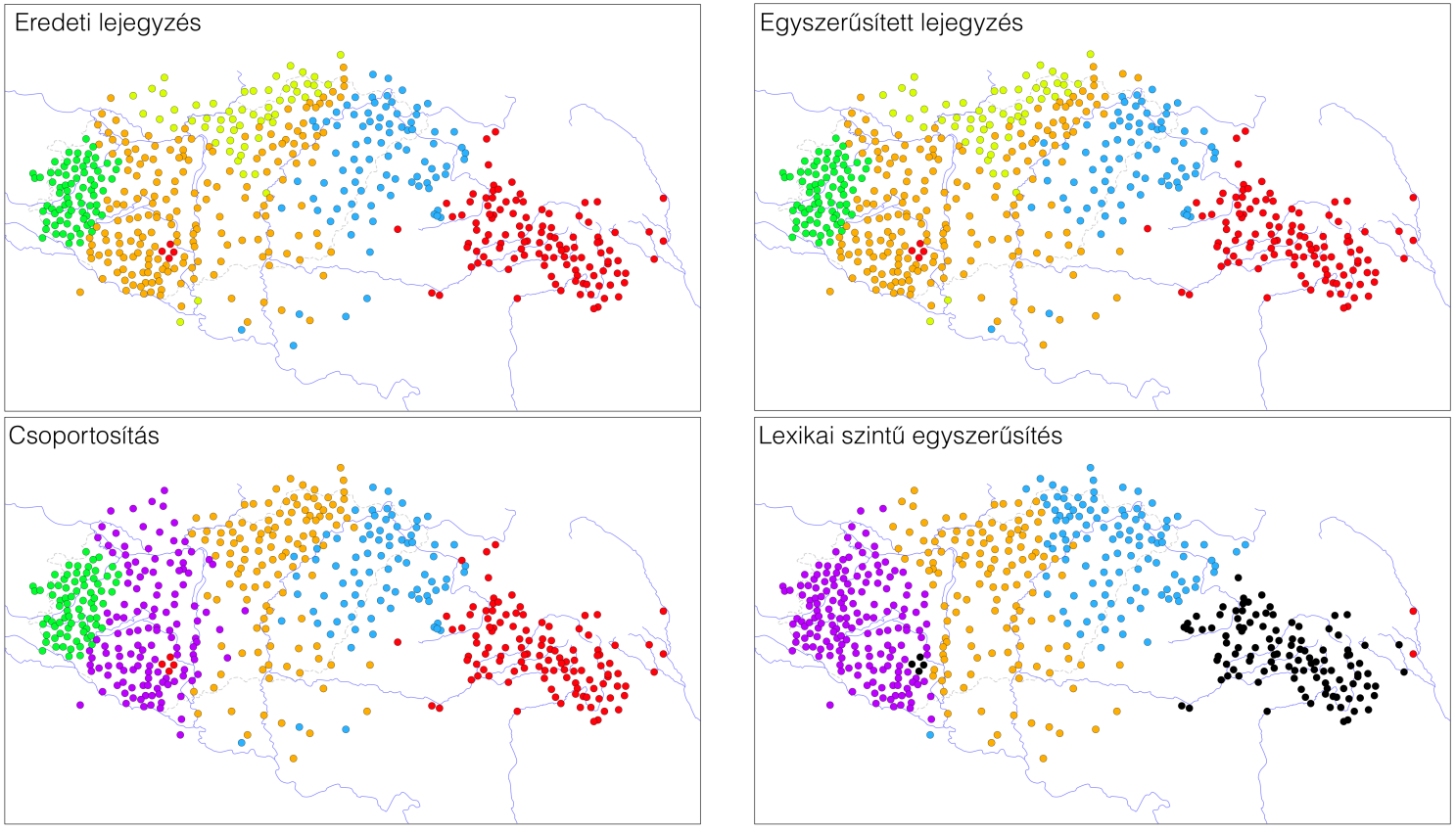
5.12. ábra: A MNyA. és a RMNyA. Ward-féle klaszteranalízises térképei különböző mátrixok alapján, 2 csoport térképezése esetén



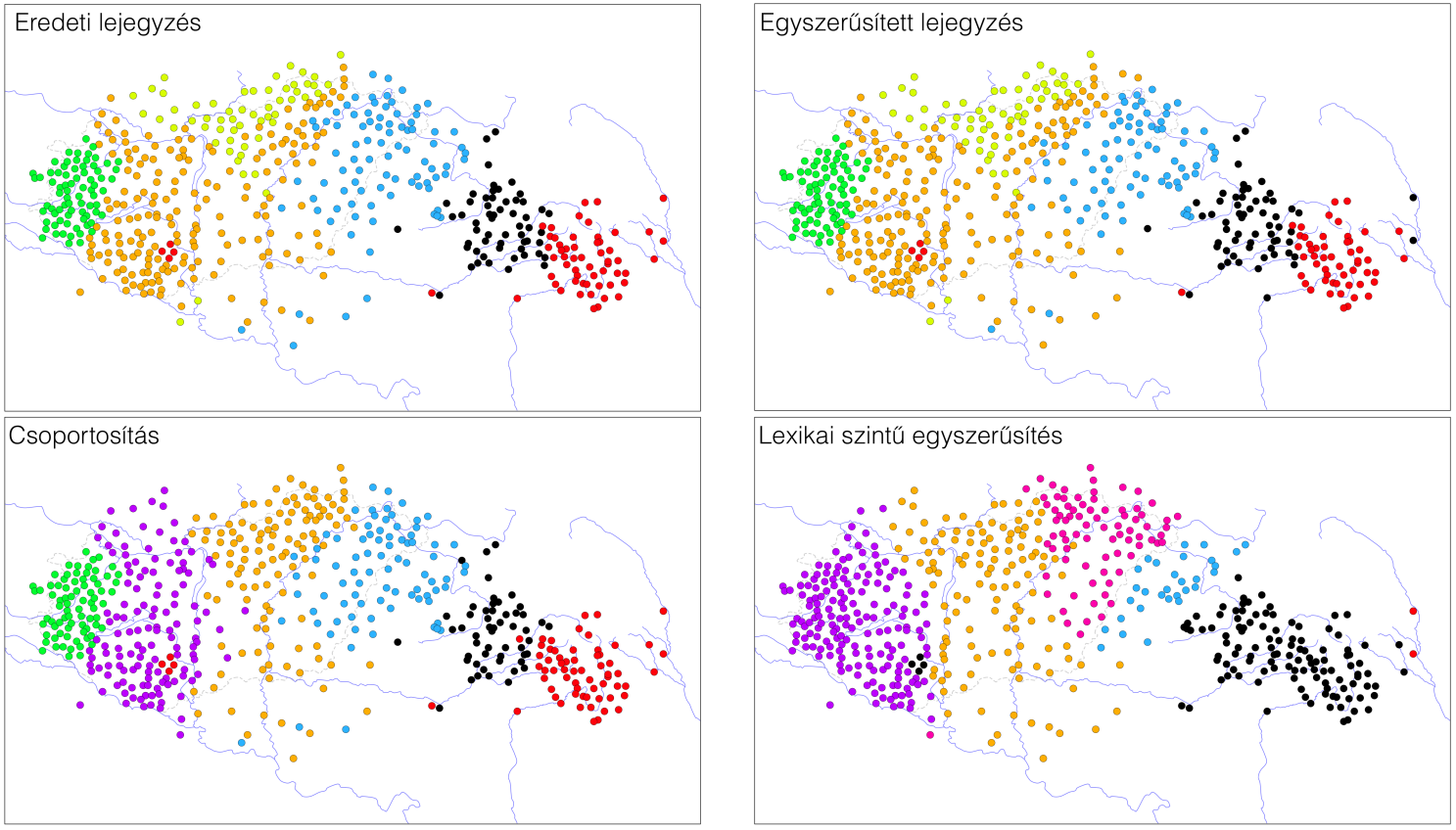
5.13. ábra: A MNyA. és a RMNyA. Ward-féle klaszteranalízises térképei különböző mátrixok alapján, 3 csoport térképezése esetén



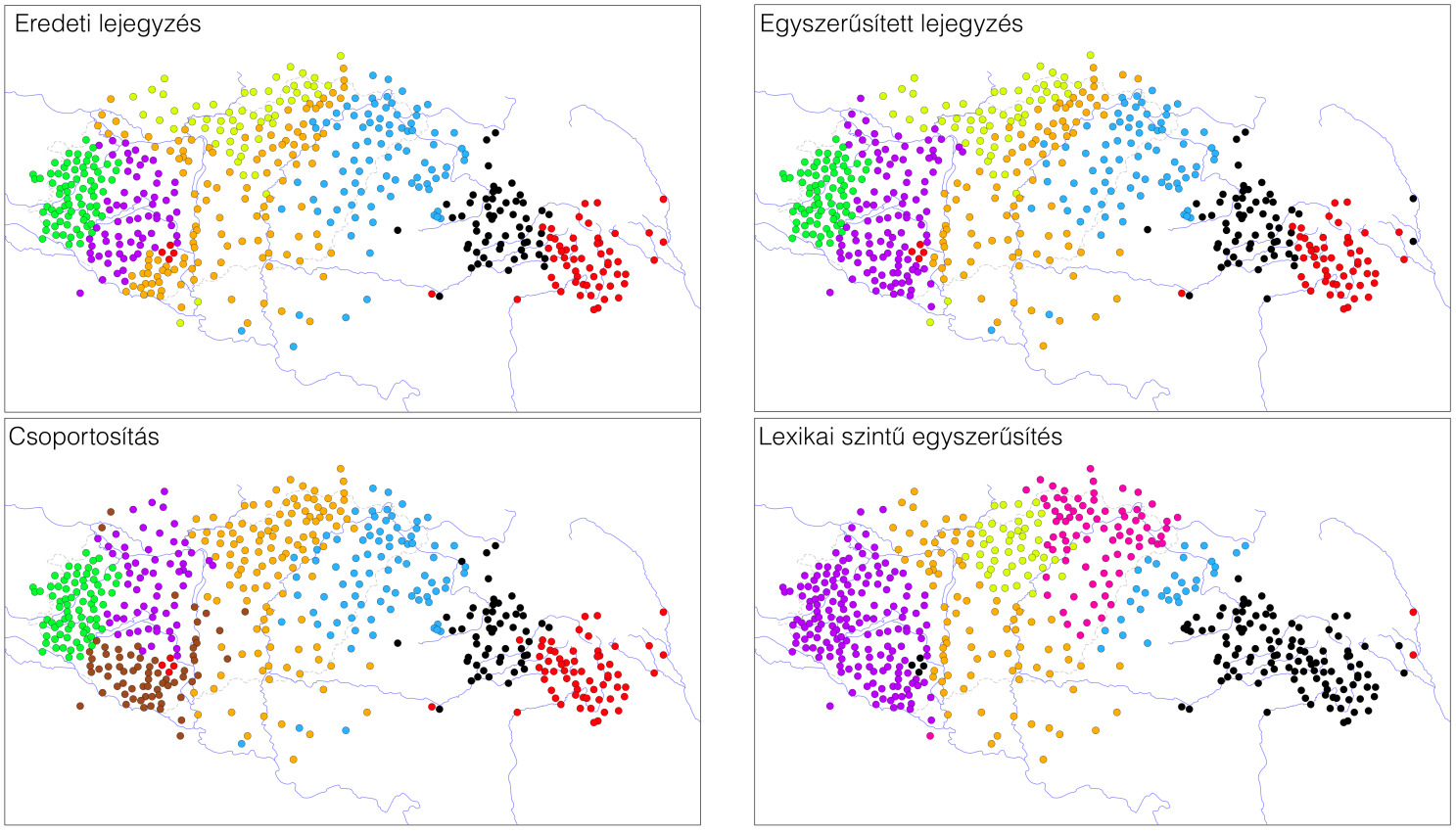
5.14. ábra: A MNyA. és a RMNyA. Ward-féle klaszteranalízises térképei különböző mátrixok alapján, 4 csoport térképezése esetén



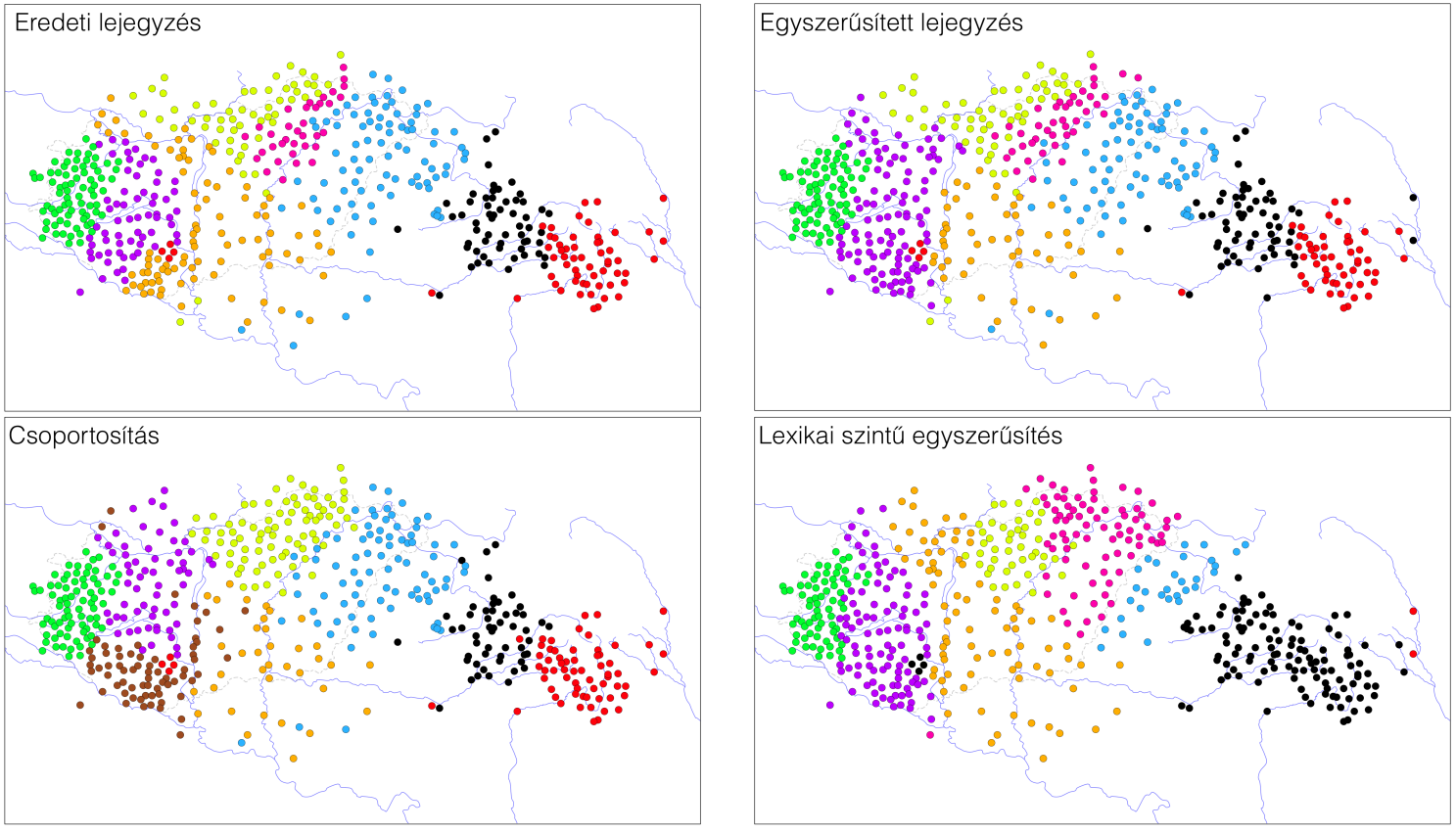
5.15. ábra: A MNyA. és a RMNyA. Ward-féle klaszteranalízises térképei különböző mátrixok alapján, 5 csoport térképezése esetén



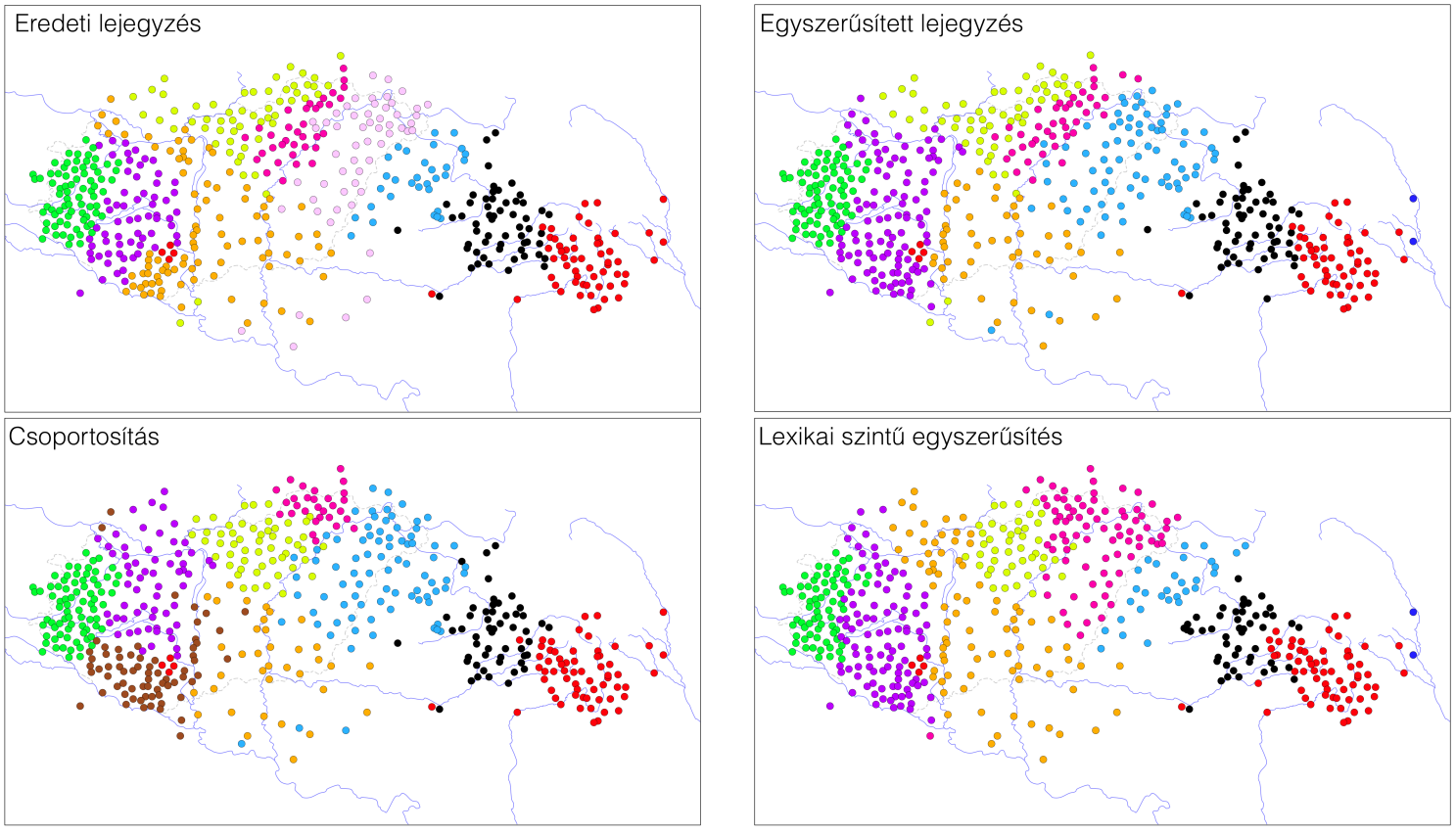
5.16. ábra: A MNyA. és a RMNyA. Ward-féle klaszteranalízises térképei különböző mátrixok alapján, 6 csoport térképezése esetén



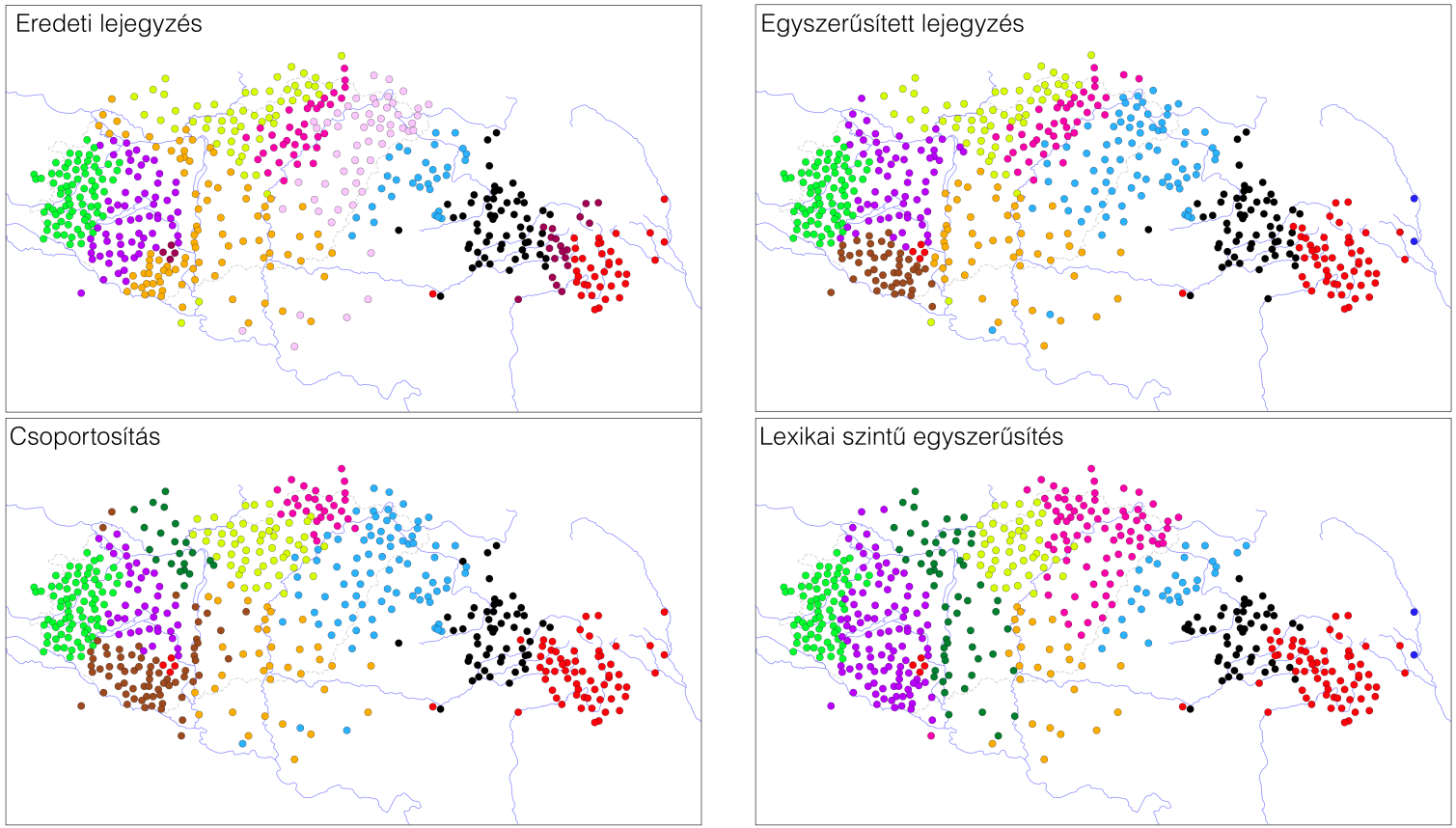
5.17. ábra: A MNyA. és a RMNyA. Ward-féle klaszteranalízises térképei különböző mátrixok alapján, 7 csoport térképezése esetén



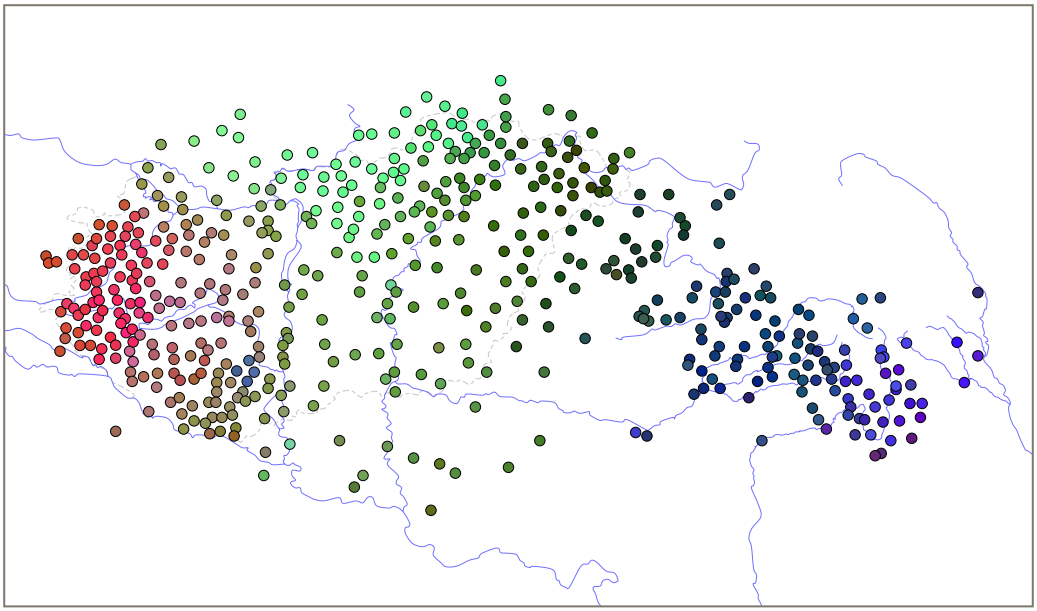
5.18. ábra: A MNyA. és a RMNyA. Ward-féle klaszteranalízises térképei különböző mátrixok alapján, 8 csoport térképezése esetén



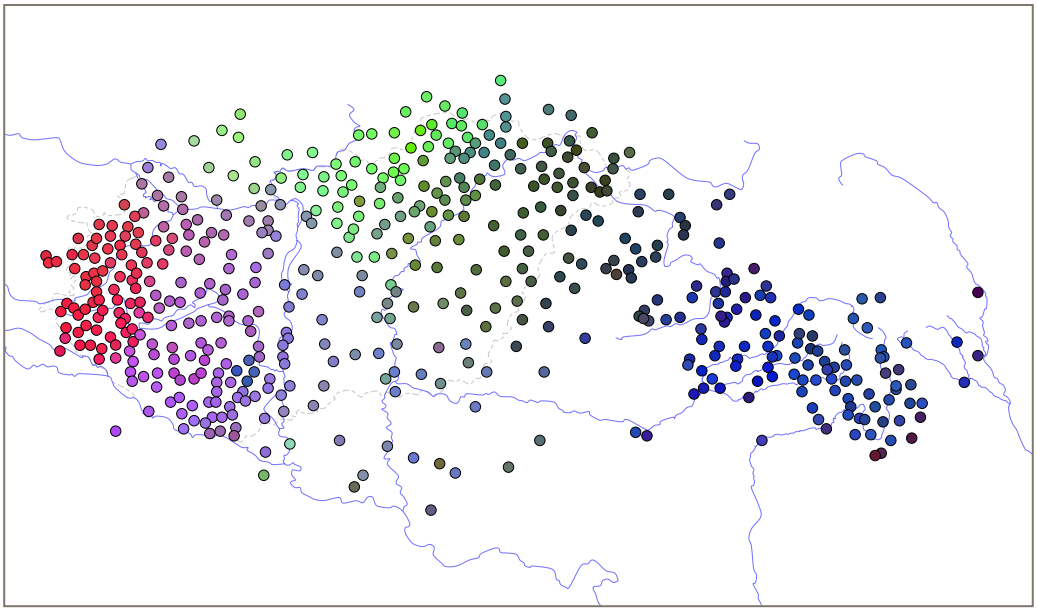
5.19. ábra: A MNyA. és a RMNyA. Ward-féle klaszteranalízises térképei különböző mátrixok alapján, 9 csoport térképezése esetén



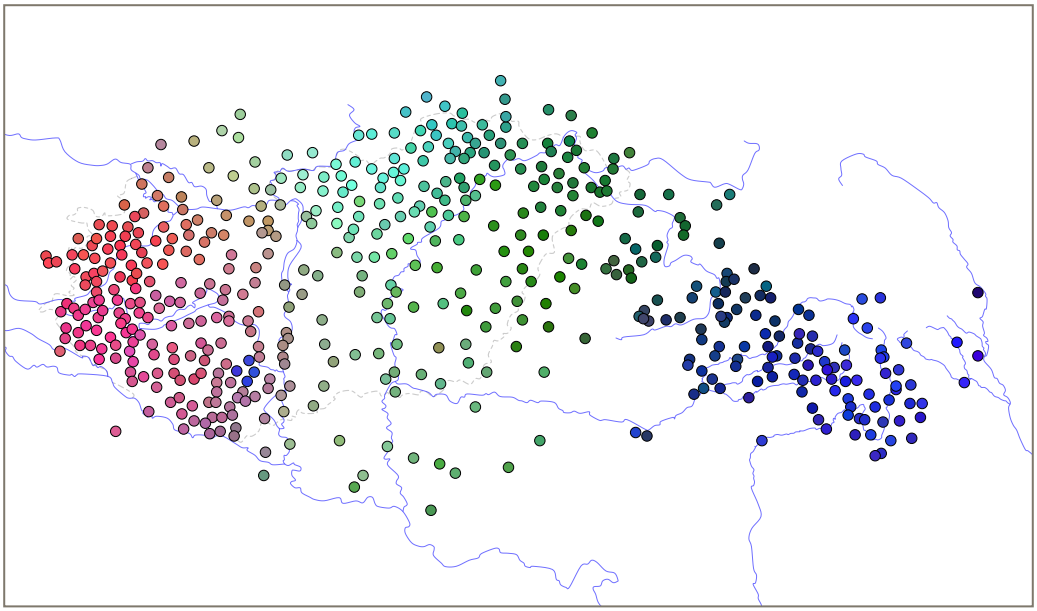
5.20. ábra: A MNyA. és a RMNyA. Ward-féle klaszteranalízises térképei különböző mátrixok alapján, 10 csoport térképezése esetén



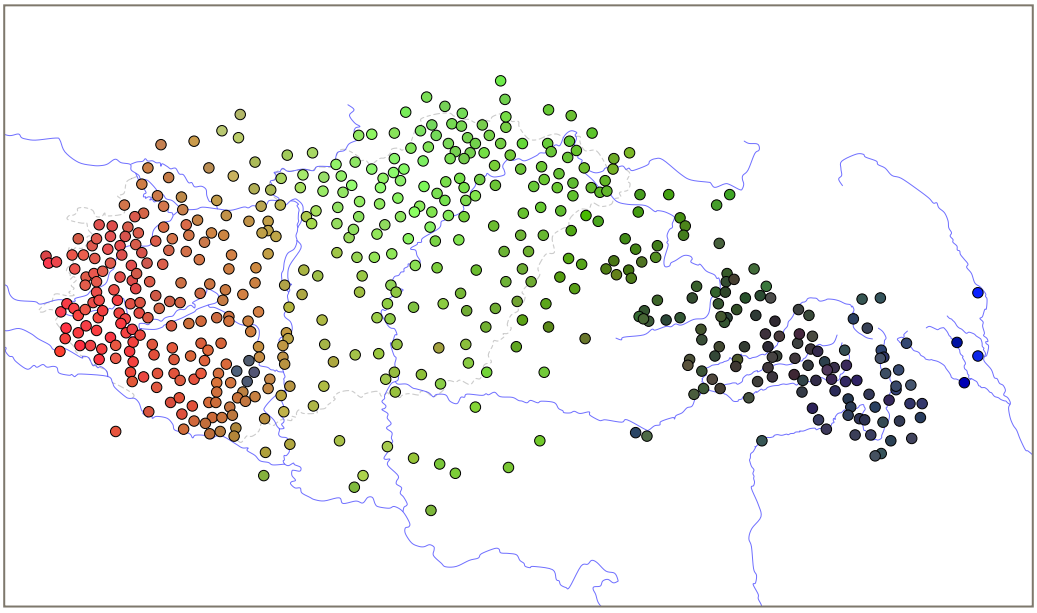
5.10. térkép: A MNyA. és a RMNyA. eredeti lejegyzése alapján készített mátrix többdimenziós skálázással készült elemzésének térképe



5.11. térkép: A MNyA. és a RMNyA. egyszerűsített, mellékjelek nélküli lejegyzése alapján készített mátrix többdimenziós skálázással készült elemzésének térképe



5.12. térkép: A MNyA. és a RMNyA. csoportosítás alapján készített mátrixa többdimenziós skálázással készült elemzésének térképe



5.13. térkép: A MNyA. és a RMNyA. lexikai szintű egyszerűsített lejegyzés alapján készített mátrix többdimenziós skálázással készült elemzésének térképe

5.7. Összefoglalás

A nyelvi hasonlósági mátrixok klaszteranalízissel és többdimenziós skálázással készült, a nyelvjárások közti határokat, illetve a nyelvi kontinuumot megmutató térképei segítségével azt jártuk körbe, hogyan használható föl a dialektometria a nyelvjárások területi felosztására, nyelvjárási határok megállapítására. A MNyA. mellett külön elemeztünk regionális atlaszokat, illetve a RMNyA. és a MCsNyA., illetve a MNyA. és a RMNyA. integrált térképlapjaiból készült mátrixokat tanulmányoztunk. Bár elemzési módszerünk objektív, az, hogy hol és milyen mértékben érzékelünk nyelvjárási törésvonalakat, számos tényező függvénye. Elsőként meghatározó a lefedett terület kiterjedése. Moldva például önálló egységnek mutatkozik a RMNyA. klasztertérképein, miként a RMNyA. és a MCsNyA. integrált elemzése alapján is, a teljes nyelvterületet lefedő mátrixok alapján azonban már nem, míg a MCsNyA. elemzésével még a moldvai kutatópontok belső nyelvjárási tagolódását is vizsgálhatjuk. Ugyanígy eltérő eredményre vezethetnek a különböző módszerrel (automatikus adatösszevetés, illetve kutatói csoportosítás), illetve az eltérő fonetikai információtartalmú lejegyzés alapján készült elemzések. Automatikus elemzés alkalmazásával biztosíthatjuk, hogy az egyes jelenségeket tényleges (atlaszbeli) előfordulási gyakoriságuknak megfelelően vesszük figyelembe. Akár a nyelvjárásszigetek eredetét, illetve az időben távoli nyelvi kapcsolatokat vizsgáljuk, akár a nyelvjárások dialektometriai elemzésen alapuló csoportosítása a célunk, kulcsfontosságúak a nagy hatókörű, az adatokban gyakran előforduló hangtani jelenségek.

6. Összegzés, kitekintés

A dialektometria a dialektológia, azon belül a nyelvföldrajz részterülete. Lényege, hogy kvantitatív megközelítésben, aggregált adatokkal dolgozik, vagyis több száz térképlap elemzésével hozza létre a kutatópontok közti nyelvi hasonlóságot vagy távolságot megmutató mátrixokat. Eszköztárát tekintve a számítógépes dialektológia része.

A dialektometriai kutatások alapvető forrásai a korábbi kutatói nemzedékek létrehozta nyelvjárási adattárak, a nyelvatlaszok, azonban nem eredeti, nyomtatott formájukban, hanem számítógépen feldolgozott, informatizált változatukban. Az informatizálás révén adataink kereshetővé, automatikusan elemezhetővé válnak, a különböző adattárak egymással integrálhatók.

Két alapvető dialektometriai módszert különböztethetünk meg az adatok összevetése szempontjából, az egyik az előzetes kutatói csoportosításon, munkatérképek létrehozásán alapuló salzburgi módszer, a másik a leginkább a groningeni egyetemhez köthető, az informatizált adatok automatikus összevetésén alapuló eljárás. A dialektometria gondolata, alapvetései már a hetvenes években megfogalmazódtak, a csoportosítás alapú kutatási módszert azonban kevesen vették át és alkalmazták a salzburgi kutatókon kívül. Az automatikus adatelemzések alkalmazása révén váltak általánosabbá a dialektometriai vizsgálatok. Ennek oka egyrészt, ahogyan magyar példákon láthattuk, hogy – ha vannak megfelelően rögzített adataink és folyamatosan fejlesztett nyelvészeti technológiánk – az automatikus összevetés szinte pillanatok alatt elvégezhető, míg az adatok csoportosítását, munkatérképek kialakítását megkívánó salzburgi módszer igen időigényes. Más nyelvekhez képest kiemelkedő mennyiségben állnak rendelkezésre informatizált magyar nyelvjárási adatok, ez az előfeltétele annak, hogy az adatok feldolgozását segítő nyelvészeti technológiák fejlesztésével és használatával viszonylag egyszerűen és hatékonyan tudjuk alkalmazni mindkét elemzési módszert, az időigényes munkafolyamatokat is optimalizálva.

A dialektometria lényege az objektivitás. A salzburgi módszer alkalmazásakor azonban maga az adatcsoportosítás, illetve az elemzendő jelenségek kiválasztása a kutatói döntések függvénye. Éppen ezért minél több munkatérképet készítünk, annál pontosabban ragadhatjuk meg a kutatópontok közti nyelvi hasonlóságot: a kutatói döntések nagy száma eredményez összességében objektívnek tekinthető eredményt, ilyenkor tehát a kvantitatív megközelítésből következik egyfajta objektivitás. Az automatikus összevetés viszont mindössze néhány száz térkép elemzése esetén is kellő mértékben objektívnek mondható. Az adatok automatikus elemzése révén mindig minden jellegzetesség hatással van az eredményekre, előfordulási gyakoriságának megfelelően. Így a nagy hatókörű, gyakrabban előforduló jelenségek (például a diftongálás vagy a gyakori magánhangzók fonetikai minősége) jobban, míg a ritkábban előfordulók kevésbé vannak hatással a kutatópontok közti nyelvi hasonlóság mértékének megállapítására. Az adatok csoportosítása indokolt lehet azonban abban az esetben, ha több adattárból származnak az adataink, hiszen ilyenkor mindenképp számolnunk kell a lejegyzői gyakorlatok közt meglévő különbségekkel, vagyis az adattárhatással. Az adattárhatás a kutatói csoportosítás alkalmazásakor kevésbé érvényesül a hasonlósági mintázatok alakulásában, mint a lejegyzés automatikus elemzésekor.

Napjainkban a dialektológiai kutatásokban már az eredeti, felvevővel rögzített hangzó formát tekinthetjük elsődlegesnek, maga a lejegyzett adat másodlagossá, folyamatosan ellenőrizhetővé, pontosíthatóvá vált. Fölmerül, van-e értelme a jövőben kialakítandó nyelvjárási korpuszoknál a fonetikai árnyalatokat is tükröző lejegyzésnek. A lejegyzés automatikus egyszerűsítése kapcsán több fejezetben is láthattunk pédát

arra, hogy a fonetikai értelemben különböző mértékben információgazdag lejegyzési szintek másféle hasonlósági mintázatokhoz vezethetnek. A dialektometriai kutatások eddigi eredményei szerint számos esetben (például nyelvjárásszigetek településtörténeti kapcsolatainak vizsgálatakor) alapvető fontosságú, hogy legyenek részletesen lejegyzett adataink. A hangtani részleteket megragadó lejegyzési stílus jövőbeni alkalmazása tehát indokolt, ez biztosíthatja a korábbi adattárakkal való összevetést is. Kutatási céljainknak megfelelően, amennyiben szükséges, automatizált módszerekkel létrehozhatjuk a lejegyzés egyszerűsített, kevesebb fonetikai információt tartalmazó formáit. Fordítva természetesen nem járhatunk el, a lejegyzésben nem rögzített információ az elemzésben felhasználhatatlan marad.

A nyelvi távolságot megmutató mátrixokat további statisztikai elemzésnek alávetve hozhatunk létre olyan térképeket, amelyek színekkel érzékeltetik, mennyire hasonlítanak egymásra az egyes kutatópontok nyelvi hasonlósági mintázatai. A klaszteranalízissel készült térképeken a kutatópontok objektív módszerekkel történő csoportosításának eredménye látható, míg többdimenziós skálázással lehetőségünk van a nyelvi kontinuum ábrázolására. A nyelvi kontinuum bemutatására alkalmas elemzéssel jobban megragadhatók a nyelvjárások tényleges hasonlósági fokozatai és a nyelvterület különböző részein meglévő párhuzamok. A kutatópontok csoportokba sorolása – noha korszerűnek tekinthető, objektív statisztikai eljárásokkal történik – mindenképpen mesterséges, több információt elrejt, mint amennyit kidomborít, és azt az egykori szemléletet tükrözi, amely a kategorizálás révén reméli a megértést. Ugyanakkor számos tanulsággal szolgálhat a későbbiekben a nyelvjárások automatikus felosztásainak és a nyelvterület MDS-térképeinek módszeres összevetése a magyar nyelvjárások korábbi (részleges vagy teljes) csoportosításaival, Balassa József 1981-es térképétől Juhász Dezső 2001-es tíz nagy nyelvjárási régiót megkülönböztető felosztásáig.

A fonetikai információt különböző mértékben tartalmazó lejegyzési szintekhez tartozó hasonlósági mátrixok klaszterezéssel és többdimenziós skálázással készült térképeinek, valamint az egyes kutatópontok – jellemzően a nyelvjárásszigetek, a több évszázados, távoli kapcsolatokkal rendelkező települések és a nyelvjárási törésvonal mentén fekvő kutatópontok – dialektometriai térképeinek összevetésével láttuk, hogy a nyelvi hasonlósági mintázatok alakulásában meghatározó szerepe van az adatokban gyakran előforduló, nagy hatókörű, hangtani jelenségeknek. Ebből kiindulva feltételezhetjük, hogy a magyar nyelvjárások közti rendszerszerű összefüggések és különbségek feltárásához a magánhangzórendszerek több szempontú elemzésével kerülhetünk még közelebb.

Bibliográfia

Aurrekoetxea, Gotzon, Karmele Fernandez-Aguirre, Jesús Rubio, Borja Ruiz, Jon Sánchez 2013. 'DiaTech': A new tool for dialectology. *Literary and Linguistic Computing* 28, 23–30.

Aurrekoetxea, Gotzon, Gotzon Santander, Iker Usobiaga, Aitor Iglesias 2016. DiaTech: tool for making dialectometry easier. *Dialectologia* 17, 1–22.

Balassa József 1891. *A magyar nyelvjárások osztályozása és jellemzése.* Budapest: MTA.

Balogh Lajos – Kiss Gábor 1992. A magyar nyelvjárások atlaszának számítógépes feldolgozása. In: Kontra Miklós (szerk.). *Társadalmi és területi változatok a magyar nyelvben.* Budapest: MTA Nyelvtudományi Intézet. 5–17.

Ben-Hur, Asa – Isabelle Guyon 2003. Detecting stable clusters using principal component analysis. In: Michael J. Brownstein – Arkady B. Khodursky (eds.). *Functional Genomics: Methods and Protocols*. Totowa: Humana Press Inc.,159–182.

Benkő Loránd 1961. Új módszerbeli lehetőségek a magyar nyelvjárástörténeti vizsgálatokban. *Magyar Nyelv* 57, 401–413.

Benkő Loránd 1967. A nyelvföldrajz történeti tanulságai. *A Magyar Tudományos Akadémia Nyelv- és Irodalomtudományok Osztályának Közleményei* 24, 29–48.

Bodó Csanád 2007a. Követéses geolingvisztikai vizsgálat Moldvában. In: Guttmann Miklós – Molnár Zoltán (szerk.). V. Dialektológiai Szimpozion. Szombathely: Berzsenyi Dániel Főiskola. 37–47.

Bodó Csanád 2007b. A moldvai magyar nyelvjárások román kölcsönszórétegének területisége. In: Benő, Attila – Fazakas Emese – Szilágyi N., Sándor (szerk.). *Nyelvek és nyelvváltozatok. Köszöntő kötet Péntek János tiszteletére.* I. Kolozsvár: Anyanyelvápolók Erdélyi Szövetsége Kiadó. 160–174.

Bodó Csanád – Vargha Fruzsina Sára 2007. *Jelenségtérképek A moldvai csángó nyelvjárás atlaszából.* CD. Budapest: ELTE Magyar Nyelvtörténeti, Szociolingvisztikai, Dialektológiai Tanszék.

Bodó Csanád – Vargha Fruzsina Sára 2008. Régi nyelvatlaszok – új módszerek. *Magyar Nyelv* 104, 335–351.

Bodó Csanád – Vargha Fruzsina Sára 2016. Román kölcsönszói hatások A moldvai csángó nyelvjárás atlasza nyelvi hasonlósági viszonyaiban. In: Czetter Ibolya, Hajba Renáta, Tóth Péter (szerk.). *VI. Dialektológiai Szimpozion.* Szombathely–Nyitra: NYME Magyar Nyelvtudományi Intézeti Tanszék, UKF Közép-európai Tanulmányok Kara, Szlovákiai Magyar Akadémiai Tanács. 167–178.

Bodó, Csanád, Fruzsina S. Vargha, Domokos Vékás 2012. Classifications of Hungarian dialects in Moldavia. In: Peti, Lehel – Vilmos Tánczsos (eds.). *Language Use, Attitudes, Strategies: Linguistic Identity and Ethnicity in the Villages of the Moldavian Csángós.* Cluj-Napoca: The Romanian Institute for Research on National Minorities. 51–69.

Borsos Balázs 2011. *A magyar népi kultúra regionális struktúrája A Magyar Néprajzi Atlasz számítógépes feldolgozása fényében* I–II. Budapest: MTA Néprajzi Intézet.

Both Csaba Attila. Megjelenés előtt. A Székelyföldi Nyelvjárási Atlasz anyagának állapota és felhasználási lehetőségei. In: Benő Attila – Gál Noémi (szerk.). A 19. Élőnyelvi Konferencia előadásai.

Browman, Catherine – Louis Goldstein 1992. Articulatory Phonology: An Overview. *Phonetica* 49, 155–180.

Chambers, Jack – Peter Trudgill 1998. *Dialectology*. 2nd edition. Cambridge: Cambridge University Press.

Deme László 1975. A magyar nyelvjárások atlaszának kérdőívei. In: Deme László – Imre Sami (szerk.). *A magyar nyelvjárások atlaszának elméleti–módszertani kérdései.* Budapest: Akadémiai Kiadó. 67–122.

Dinkin, Aaron – Keelan Evanini 2010. An Eleméntàry Linguistic Definition of Upstate New York. *Penn Working Papers in Linguistics* 16, 36–45.

Fazakas (Gál) Noémi 2013. Az *l* kiesése mint nyelvi változó a Vöő István-féle hangoskönyvben. In: Kontra Miklós – Németh Miklós – Sinkovics Balázs (szerk.). *Elmélet és empíria a szociolingvisztikában.* Budapest: Gondolat Kiadó. 402–416.

N. Fodor János 2011. A kórógyi nyelvjárás nyelvföldrajzi tanulságai. In: Szoták Szilvia (szerk.). *Magyar nyelv és kultúra a Kárpát-medencében.* Dunaszerdahely: Gramma Nyelvi Iroda. 62–73.

N. Fodor János 2012. A magyar magánhangzórendszert érintő hangeltolódás tendenciájának nyelvjárási vonatkozásai.: Az északkeleti nyitódó kettőshangzók kialakulásának lehetséges módjáról. In: É. Kiss Katalin, Hegedűs Attila (szerk.), *Nyelvelmélet és dialektológia 2*. Piliscsaba: PPKE BTK. 136–153.

Fodor Katalin 2001. A nyelvjárási hangtani jelenségek. In: Kiss Jenő (szerk.). *Magyar dialektológia*. Budapest: Osiris Kiadó. 325–350.

Goebl, Hans 2002. Analyse dialectométrique des structures de profondeur de l'ALF. *Revue de linguistique romane* 66, 5–63.

Goebl, Hans 2005. La dialectométrie corrélative. Un nouvel outil pour l'étude de l'aménagement dialectal de l'espace par l'homme. *Revue de linguistique romane* 69. 321–367.

Goebl, Hans. 2006. Recent advances in Salzburg dialectometry. *Literary and Linguistic Computing* 21. 411–435.

Goebl, Hans 2008. La dialettometrizzazione integrale dell'AIS. Prezentazione dei primi risultati. *Revue de Linguistique Romane* 72, 25–113.

Goebl, Hans 2010. Dialectometry and quantitative mapping. In: Alfred Lameli, Roland Kehrein, Stefan Rabanus (eds.). *Language and Space. Vol. 2: Language Mapping.* Berlin: De Gruyter Mouton. 433–457.

Goebl, Hans 2011. Introduction aux problèmes et méthodes de l'«École dialectométrique de Salzbourg» (avec des exemples gallo-, italo- et ibéroromans). In: Afonso Álvarez Pérez, Ernestina Carrilho, Catarina Magro (eds.). *Proceedings of the International Symposium on Limits and Areas in Dialectology (LimiAr), Lisbon 2011*. Lisboa: Centro de Linguística da Universidade de Lisboa. 117–166

Gooskens, Charlotte 2005. Traveling time as a predictor of linguistic distance. *Dialectologia et Geolinguistica* 13, 38-42.

Grieve, Jack 2014. A comparison of statistical methods for the aggregation of regional linguistic variation. In Szmrecsanyi, Benedikt – Bernhard Wälchli (eds.). *Aggregating Dialectology, Typology, and Register Analysis. Linguistic Variation in Text and Speech.* Berlin: De Gruyter.

Heeringa, Wilbert 2004. *Measuring Dialect Pronunciation Differences using Levenshtein Distance.* Groningen Dissertations is Linguistics. Groningen: University of Groningen.

Heeringa, Wilbert – Frans Hinskens 2011. The Measurement of Dutch Dialect Change: Lexicon versus Morphology versus Sound Components. *Taal en Tongval* 63, 79–98.

Heeringa, Wilbert – John Nerbonne 2001. Computational Comparison and Classification of Dialects. *Dialectologia et Geolinguistica* 9, 69–83.

Heeringa, Wilbert – John Nerbonne 2013. Dialectometry. In: Frans Hinskens & Johan Taeldeman (eds.), *Language and Space. An International Handbook of Linguistic Variation, Volume III: Dutch.* (Handbook of Linguistics and Communication Science (HSK) 30/3). Berlin & New: York Walter de Gruyter. 624–646.

Hegedűs Attila 2012. *A vonzatosság a magyar nyelvjárásokban.* Budapest–Piliscsaba: Szent István Társulat.

Iglai Edit 2016. Nyelvjárásközi egybevető vizsgálat A magyar nyelvjárások atlasza és az Új magyar nyelvjárási atlasz alapján. Doktori disszertáció. ELTE BTK, Nyelvtudományi Doktori Iskola.

Imre Samu 1971. *A mai magyar nyelvjárások rendszere*. Budapest: Akadémiai Kiadó.

Juhász Dezső 2001a. A nyelvföldrajz. In: Kiss Jenő (szerk.). *Magyar dialektológia*. Budapest: Osiris Kiadó. 92–110.

Juhász Dezső 2001b. A magyar nyelvjárások területi egységei. In: Kiss Jenő (szerk.). *Magyar dialektológia*. Budapest: Osiris Kiadó. 262–315, 460–461.

Juhász Dezső 2011. A magyar nyitódó kettőshangzók történetéről a tér és idő dimenziójában. In: Bakró-Nagy Marianne – Forgács Tamás (szerk.). *A nyelvtörténeti kutatások újabb eredményei VI.* Szeged: Szegedi Tudományegyetem, Magyar Nyelvészeti Tanszék. 123–128.

Juhász Dezső 2007. A nyelvföldrajz történeti tanulságai – egy klasszikus tanulmány negyven év távlatából. In: Guttmann Miklós – Molnár Zoltán (szerk.). *V. Dialektológiai Szimpozion.* Szombathely: Berzsenyi Dániel Főiskola. 133–138.

Király Lajos 1990. *Nyelvjárási jelenséghatárok a Dél-Dunántúlon.* MNyTK. 186. Budapest: Magyar Nyelvtudományi Társaság.

Kiss Jenő 2001a. A nyelvjárási tudat. In: Kiss Jenő (szerk.). *Magyar dialektológia*. Budapest: Osiris Kiadó. 210–217.

Kiss Jenő 2001b. A nyelvjárások osztályozása. In: Kiss Jenő (szerk.). *Magyar dialektológia*. Budapest: Osiris Kiadó. 72–81.

Kocsis Zsuzsanna – Vargha Fruzsina Sára 2016. A magyar nyelvjárások atlasza és A romániai magyar nyelvjárások atlasza integrált dialektometriai elemzése. In: Czetter Ibolya, Hajba Renáta, Tóth Péter (szerk.). *VI. Dialektológiai Szimpozion.* Szombathely–Nyitra: NYME Magyar Nyelvtudományi Intézeti Tanszék, UKF Közép-európai Tanulmányok Kara, Szlovákiai Magyar Akadémiai Tanács. 193–207.

Kontra Miklós 2003. A szép magyar beszéd és a csúnya. In: Kontra Miklós (szerk.). *Nyelv és társadalom a rendszerváltáskori Magyarországon*. Budapest: Osiris Kiadó. 240–255.

Kontra Miklós, Németh Miklós, Sinkovics Balázs 2016. Szeged nyelve a 21. században. Budapest: Gondolat.

Labov, William, Sharon Ash, Charles Boberg 2006. *The Atlas of North American English. Phonetics, Phonology and Sound Change.* Berlin/New York: Mouton de Gruyter.

P. Lakatos Ilona, T. Károlyi Margit, Iglai Edit 2012. *Változó nyelvhasználat a hármas határ mentén. Többdimenziós nyelvföldrajzi térképlapok tanúságai.* Budapest: Tinta Könyvkiadó.

Leemann, Adrian, Maria-José Kolly, Ross Purves, David Britain, Elvira Glaser (2016) Crowdsourcing Language Change with Smartphone Applications. PLoS ONE 11(1): e0143060

Levshina, Natalia 2015. *How to do Liguistics with R. Data exploration and statistical analysis*. Amsterdam / Philadelphia: John Benjamins Publishing Company.

Lőrincze Lajos 1975. Az anyaggyűjtés módszere. In: Deme László – Imre Samu (szerk.). *A magyar nyelvjárások atlaszának elméleti–módszertani kérdései.* Budapest: Akadémiai Kiadó. 167–204

Mathussek, Andrea 2016. On the problem of field worker isoglosses. In: Marie-Hélène Côté, Remco Knooihuizen & John Nerbonne (eds.). *The future of dialects,* Berlin: Language Science Press. 99–116.

MCsNyA. = *A moldvai csángó nyelvjárás atlasza 1–2.* Gálffy Mózes – Márton Gyula – Szabó T. Attila (szerk.) 1991. (A kiadás előkészítői: Murádin László – Péntek János.) MNyTK. 193. Budapest: Magyar Nyelv- tudományi Társaság.

Menyhárt József – Presinszky Károly 2013. A szórványos *ö*-zés vizsgálata a Csallóközben generációk és iskolázottság szerint. In: Kontra Miklós – Németh Miklós – Sinkovics Balázs (szerk.). *Elmélet és empíria a szociolingvisztikában.* Budapest: Gondolat Kiadó. 444–458.

MNyA. = *A magyar nyelvjárások atlasza 1–6.* Deme László – Imre Samu (szerk.). 1968–1977. Budapest: Akadémiai Kiadó.

Montemagni, Simonetta. 2008. The space of Tuscan dialectal variation. A correlation study. *International Journal of Humanities and Arts Computing* 2, 135–152.

Cs. Nagy Lajos 2011. Úton a Háromszéki nyelvatlasz informatizálása felé. In: Bárth M. János – Vargha Fruzsina Sára (szerk.). Hangok – helyek. Budapest: ELTE Magyar Nyelvtudományi és Finnugor Intézet. 61–73.

Navarro, Gonzalo 2001. A guided tour to approximate string matching. *ACM computing surveys (SUR)* 33, 31-88

Nerbonne, John, Wilbert Heeringa, Eric van den Hout, Peter van de Kooi, Simone Otten, Willem van de Vis. 1996. Phonetic distance between Dutch dialects. In: Gert Durieux, Walter Daelemans, Steven Gillis (eds.). *CLIN VI: Papers from the Sixth CLIN Meeting.* Antwerp*:* Centre for Dutch Language and Speech (UIA). 185–202.

Nerbonne, John – Peter Kleiweg 2003. Lexical distance in LAMSAS. In: John Nerbonne – William Kretzschmar (eds.). *Computational Methods in Dialectometry. Special issue of Computers and the Humanities*, 37. 339–357.

Nerbonne, John – William A. Kretzschmar, Jr 2013. Dialectometry++. *Literary and Linguistic Computing* 28, 2–12.

Péntek János 2005. Magyar nyelv- és nyelvjárásszigetek Romániában. *Magyar Nyelv* 101, 406–413.

Péntek János 2014. A moldvai magyarokról és a csángó elnevezésről. *Magyar Nyelv* 110, 406–416.

Perea, Maria-Pilar 2010. Catalan geolinguistics and new technical procedures. *Dialectologia, Geolinguistics around the World,* Special Issue I, 147–160.

Pickl, Simon, Aaron Spettl, Simon Pröll, Stephan Elspaß, Werner König, Volker Schmidt 2014. Linguistic distances in dialectometric intensity estimation. *Journal of Linguistic Geography* 2. 25–40.

Pickl, Simon 2016. Fuzzy dialect areas and prototype theory: Discovering latent patterns in geolinguistic variation. In: Marie-Hélène Côté, Remco Knooihuizen, John Nerbonne (eds.). *The future of dialects. Selected papers from Methods in Dialectology XV.* Berlin: Language Science Press. 75–97.

Prokic, Jenna – John Nerbonne 2008. Recognizing groups among dialects. *International Journal of Humanities and Arts Computing* 1, 153–172.

RMNyA. = *A romániai magyar nyelvjárások atlasza*. I–XI. kötet. Murádin László (gyűjt.) – Juhász Dezső (szerk.), 1995–2010. Budapest: Magyar Nyelvtudományi Társaság.

Sándor Anna 2007. Kísérlet a palóc nyelvjárások nyugati határainak pontosítására. In: Zelliger Erzsébet (szerk.). *Nyelv, területiség, társadalom. A 14. Élőnyelvi Konferencia (Bük, 2006. október 9–11.) előadási.* MNyTK. 228. Budapest: Magyar nyelvtudományi Társaság. 45–54.

Séguy, Jean 1973 La dialectométrie dans l'Atlas linguistique de la Gascogne. *Revue de Linguistique Romane* 37, 1–24.

S–ZA. = *Somogy–zalai nyelvatlasz.* Király Lajos 2005. Budapest: Magyar Nyelvtudományi Társaság.

Szabó József 1990. Magyarországi és jugoszláviai magyar nyelvjárásszigetek. Dél-Alföldi Évszázadok 3. Békéscsaba–Kecskemét–Szeged: Csongrád Megyei Levéltár.

Valls, Esteve, John Nerbonne, Jelena Prokic, Martijn Wieling, Esteve Clua & Maria-Rosa Lloret 2012. Applying the Levenshtein Distance to Catalan dialects: A brief comparison of two dialectometric approaches. *Verba. Annuario Galego de Filoloxía* 39: 35–61.

Vargha Fruzsina Sára 2007a. Állatok kicsinyeinek megnevezése a keleti magyar nyelvjárásokban. In: Hoffmann István – Juhász Dezső (szerk.). *Nyelvi identitás és a nyelv dimenziói.* Debrecen-Budapest: Nemzetközi Magyarságtudományi Társaság. 237–248.

Vargha Fruzsina Sára 2007b. Nyelvi változók A magyar nyelvjárások atlasza hangfelvételeiben. In: Guttmann Miklós – Molnár Zoltán (szerk.). *V. Dialektológiai Szimpozion*. Szombathely: Berzsenyi Dániel Főiskola. 279–289.

Vargha Fruzsina Sára 2011. Beszélő térképlapok A magyar nyelvjárások atlaszából. In: Bárth M. János – Vargha Fruzsina Sára (szerk.). *Hangok – helyek*. Budapest: ELTE Magyar Nyelvtudományi és Finnugor Intézet.

Vargha Fruzsina Sára 2013. A hangzó adat szerepe a magyar dialektológiában. In: Szoták Szilvia – Vargha Fruzsina Sára (szerk.). *Változó nyelv, nyelvváltozatok, területiség: A VII. Hungarológiai Kongresszus nyelvészeti tanulmányai.* Kolozsvár: Egyetemi Műhely Kiadó–Bolyai Társaság. 94–204.

Vargha Fruzsina Sára 2013. Nyelvatlaszok a szekrényben. In: Benő Attila, Fazakas Emese, Kádár Edit (szerk.). *„…hogy legyen a víznek lefolyása…": Köszöntő kötet Szilágyi N. Sándor tiszteletére.* Kolozsvár: Erdélyi Múzeum-Egyesület. 523–529.

Vargha Fruzsina Sára 2015a. Lexikai, fonológiai, fonetikai stabilitás (és relevancia) a magyar nyelvjárásokban. In: É. Kiss Katalin – Hegedűs Attila – Pintér Lilla (szerk.). *Nyelvelmélet és dialektológia 3.* Budapest–Piliscsaba: PPKE BTK Elméleti Nyelvészeti Tanszék – Magyar Nyelvészeti Tanszék. 243–261.

Vargha Fruzsina Sára 2015b. Atlaszintegrálás és kvantitatív adatelemzés. In: Bárth János – Bodó Csanád – Kocsis Zsuzsanna (szerk.). *A nyelv dimenziói. Tanulmányok Juhász Dezső tiszteletére.* Budapest: ELTE BTK. 242–249.

Vargha, Fruzsina S. Megjelenés előtt. The Impact of Phonetic Information in Dialectometry – a Case Study of Hungarian Dialect Atlases. *Dialectologia*.

Vargha, Fruzsina S. 2016a. Linguistic self-hatred and distance from standard Hungarian. Előadás. Sociolinguistics Symposium 21. Murcia, June 15. http://frufru.web.elte.hu/eloadasok/SS21__varghafru.pdf

Vargha Fruzsina Sára 2016b. A romániai magyar nyelvjárások atlasza informatizált térképlapjainak kvantitatív nyelvföldrajzi vizsgálata. *Magyar Nyelv* 112. 152–163.

Vargha Fruzsina Sára – Vékás Domokos 2009. Magyar nyelvjárási adattárak vizsgálata interaktív dialektometriai térképekkel. Előadás. Magyar Nyelvtudományi Társaság felolvasóülése, 2009. március 24. http://www.bihalbocs.hu/eloadas/dialektometria_20090324.pdf

Vékás Domokos 2007. Számítógépes dialektológia. In: Guttmann Miklós – Molnár Zoltán (szerk.). *V. Dialektológiai Szimpozion.* Szombathely: Berzsenyi Dániel Főiskola. 289–293.

Zelliger Erzsébet 1988. Településtörténeti kérdések a szóföldrajz tükrében. In: Kiss Jenő – Szűcs László (szerk.). A magyar nyelv rétegződése 1–2. Budapest: Akadémiai Kiadó. 2: 1029–1040.

Adatelemzés, térképezés és statisztika

Bihalbocs: magyar nyelvjárási lejegyző, adatbáziskezelő, térképező és elemző program. Fejleszti: Vékás Domokos (1996–) és Vargha Fruzsina Sára (2005–). http://www.bihalbocs.hu/.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

ade4 package (R): Dray, S. and Dufour, A.B. (2007): The ade4 package: implementing the duality diagram for ecologists. Journal of Statistical Software. 22(4): 1-20.